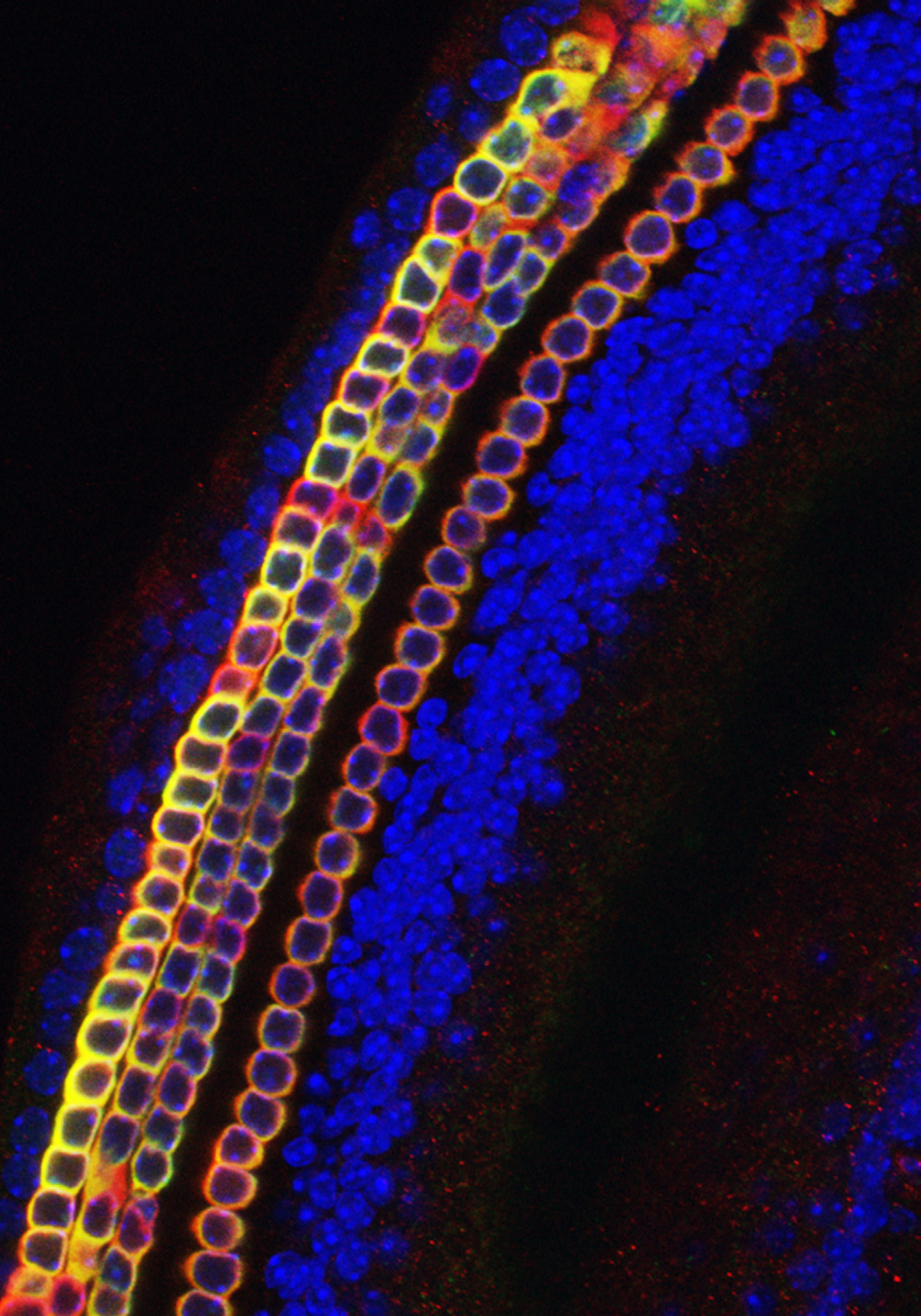
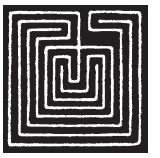
Dædalus

Journal of the American Academy of Arts & Sciences

Winter 2015

What is the Brain Good For?





Inside front cover : The receptor for auditory stimuli, the organ of Corti, includes a single row of inner hair cells that convey information about sound to the brain. The three adjacent rows of outer hair cells constitute a mechanical amplifier that increases the ear's sensitivity and sharpens its frequency tuning. In this confocal micrograph of the mouse's cochlea, both green and red labels distinguish the hair cells. The nuclei of all the cells, including the arcs of cells bracketing the hair cells, are marked with a blue dye.

Dædalus is designed by Alvin Eisenman.

Dædalus

Journal of the American Academy of Arts & Sciences

The labyrinth designed by Daedalus for King Minos of Crete, on a silver tetradrachma from Cnossos, Crete, c. 350–300 B.C. (35 mm, Cabinet des Médailles, Bibliothèque Nationale, Paris). "Such was the work, so intricate the place, / That scarce the workman all its turns cou'd trace; / And Daedalus was puzzled how to find / The secret ways of what himself design'd." – Ovid, *Metamorphoses*, Book 8

Dædalus was founded in 1955 and established as a quarterly in 1958. The journal's namesake was renowned in ancient Greece as an inventor, scientist, and unriddler of riddles. Its emblem, a maze seen from above, symbolizes the aspiration of its founders to "lift each of us above his cell in the labyrinth of learning in order that he may see the entire structure as if from above, where each separate part loses its comfortable separateness."

The American Academy of Arts & Sciences, like its journal, brings together distinguished individuals from every field of human endeavor. It was chartered in 1780 as a forum "to cultivate every art and science which may tend to advance the interest, honour, dignity, and happiness of a free, independent, and virtuous people." Now in its third century, the Academy, with its nearly five thousand elected members, continues to provide intellectual leadership to meet the critical challenges facing our world.

Dædalus Winter 2015
Issued as Volume 144, Number 1

Editorial offices: *Dædalus*, American Academy of Arts & Sciences, 136 Irving Street, Cambridge MA 02138. Phone: 617 576 5085. Fax: 617 576 5088. Email: daedalus@amacad.org.

Library of Congress Catalog No. 12-30299.

Dædalus (ISSN 0011-5266; E-ISSN 1548-6192) is published quarterly (winter, spring, summer, fall) by The MIT Press, One Rogers Street, Cambridge MA 02142-1209, for the American Academy of Arts & Sciences. An electronic full-text version of *Dædalus* is available from The MIT Press. Subscription and address changes should be addressed to MIT Press Journals Customer Service, One Rogers Street, Cambridge MA 02142-1209. Phone: 617 253 2889; U.S./Canada 800 207 8354. Fax: 617 577 1545. Email: journals-cs@mit.edu.

Printed in the United States of America by Cadmus Professional Communications, Science Press Division, 300 West Chestnut Street, Ephrata PA 17522.

Newsstand distribution by Ingram Periodicals Inc., 18 Ingram Blvd., La Vergne TN 37086.

Postmaster: Send address changes to *Dædalus*, One Rogers Street, Cambridge MA 02142-1209. Periodicals postage paid at Boston MA and at additional mailing offices.

Subscription rates: Electronic only for nonmember individuals – $47; institutions – $129.

Canadians add 5% GST. Print and electronic for nonmember individuals – $52; institutions – $144. Canadians add 5% GST. Outside the United States and Canada add $23 for postage and handling. Prices subject to change without notice. Institutional subscriptions are on a volume-year basis. All other subscriptions begin with the next available issue.

Single issues: $14 for individuals; $36 for institutions. Outside the United States and Canada add $6 per issue for postage and handling. Prices subject to change without notice.

Claims for missing issues will be honored free of charge if made within three months of the publication date of the issue. Claims may be submitted to journals-cs@mit.edu. Members of the American Academy please direct all questions and claims to daedalus@amacad.org.

Advertising and mailing-list inquiries may be addressed to Marketing Department, MIT Press Journals, One Rogers Street, Cambridge MA 02142-1209. Phone: 617 253 2866. Fax: 617 253 1709. Email: journals-info@mit.edu.

The typeface is Cycles, designed by Sumner Stone at the Stone Type Foundry of Guinda CA. Each size of Cycles has been separately designed in the tradition of metal types.

Neuroscience: The Study of the Nervous System & Its Functions

Fred H. Gage

> Any man could, if he were so inclined,
> be the sculptor of his own brain.
>
> – Santiago Ramón y Cajal,
> *Advice for a Young Investigator* (1897)

Neuroscience is the scientific study of the nervous system (the brain, spinal cord, and peripheral nervous system) and its functions. The belief that the brain is the organ that controls behavior has ancient roots, dating to early civilizations that connected loss of function to damage to parts of the brain and spinal cord. But the modern era of neuroscience began – and continues to progress – with the development of tools, techniques, and methods used to measure in ever more detail and complexity the structure and function of the nervous system. The modern era of neuroscience can be traced to the 1890s, when the Spanish pathologist Santiago Ramón y Cajal used a method developed by the Italian physician Camillo Golgi to stain nerve tissues to visualize the morphology and structure of the neurons and their connections. The detailed description of the neurons and their connections by Cajal, his students, and their followers led to the "neuron doctrine," which proposed that the neuron is the functional unit of the nervous system.

We now know that the human brain contains approximately one hundred billion neurons and that these neurons have some one hundred trillion connections, forming functional and definable circuits. These neural circuits can be organized into larger

FRED H. GAGE, a Fellow of the American Academy since 2005, is Professor in the Laboratory of Genetics and the Vi and John Adler Chair for Research on Age-Related Neurodegenerative Disease at the Salk Institute for Biological Studies. He studies the unanticipated plasticity and complexity represented in the brain.

doi:10.1162/DAED_e_00313

networks and anatomical structures that integrate information across and between all sensory modalities – including hearing, seeing, touching, tasting, and smelling – from all parts of the nervous system. These networks process information derived from the internal and external environment, and the consequence of processing this sensory information is *cognition*, a concept that includes learning and memory, perception, sleep, decision-making, emotions, and all forms of higher information processing. In response to a simple or complex sensory experience, an organism responds or behaves. The behavior can be simple, like a motor reflex in response to pain, or more complicated, like playing squash, working a crossword puzzle, or painting. However, behavior is not just what an organism *does* in response to a stimulus or sensory input; it is most often what an organism *chooses* to do from a variety of available options in response to a complex set of environmental conditions. Thus, except for rare responses, like simple reflexes, a behavior is expressed in response to a combination of the immediate sensory stimuli integrated over time with cognition.

Neuroscientists conduct experiments to understand how sensory information is processed to lead to behavior. Because of the obvious complexity of the brain, neuroscientists conduct their studies at different levels of depth. While *neurons* are conceivably the smallest units in which behavior can be clearly described, the neuron is itself made up of unique anatomical features, including a *soma* (cell body), *dendrites* (the antennae branching from the soma that receive signals from other neurons), and *axons* (the processes extending from the soma that send signals to other neurons).

These neuronal components in turn contain subcellular specializations that represent the defining features of the neuron. Key among these specializations is the *synapse*: a structure shared by the dendrite and the axon that represents the junction point for the principal form of communication between two neurons. On the dendritic side of the synapse is a structure called a *spine*, which responds to signals from the axon. On the axonal side is the *bouton*, which has vesicles containing neurotransmitters – the signals to which the spine responds. Each neuron can have multiple dendrites and thousands of spines connected to comparable numbers of boutons, which together form the thousands of synapses that make up the units of communication between individual neurons.

In the soma, specialized proteins and microstructures form the basis for the intracellular communications and physiological features of neurons; for example, specialized enzymes produce the neurotransmitters and vesicles that are used in the bouton to signal the spine. Furthermore, specialized cytoskeletal proteins form long and active extensions that allow the dendrites and axons to act as a supply train for the vesicles and neurotransmitters that are made in the soma and transported to the boutons. Among the most important proteins in the neuron are those that form the ion channels. These are multi-protein structures that span the neuron's membrane and allow neurons to form electrochemical gradients, which are the driving forces of activity in neurons.

These proteins – which are crucial for the functioning of a neuron – are all the products of genes that are the functional unit of the genome, which is located in the nucleus of the neuron. Each neuron's genome contains about twenty thousand genes, but different genes are expressed in different types of neurons, and it is this unique expression pattern of genes in any particular neuron that provides its unique identity.

Even from this brief survey of the different levels of brain connectivity it is clear that it would be impossible to study the total functioning of the brain – from behavior to gene expression – in one experiment. So neuroscientists instead generally choose some limited number of brain-activity levels to probe as they address their own specific questions. The many methods used to study the nervous system differ depending on the level of analysis, but they fall generally into one of two categories: *descriptive*, for generating hypotheses, or *manipulative*, for testing hypotheses.

One type of descriptive study is a case study, in which an experimenter observes the behavior of a person or group of people before, during, or after an event that may demonstrate a role for the nervous system. The circumstances surrounding the event are usually non-repeatable and cannot be precisely reconstructed in a laboratory setting. One could argue that these are not true experiments, but these studies have revealed substantial information about aspects of neural function that was previously unknown. One remarkable example is the case of H.M., a patient whose epilepsy was treated through removal of a portion of his brain called the hippocampus and parts of the temporal lobe on both sides of his brain. As a result of the surgery, which did successfully control his epilepsy, he displayed a unique form of memory loss, and his behavior was examined over a period of forty years from the time of his operation until he died, revealing through careful documentation and experimentation some of the most important concepts about human learning and memory. Another important case study is that of Phineas Gage, a railroad worker involved in an accident in 1848 that resulted in an iron rod passing through his skull. The rod entered the left side of his head, passing just behind his left eye, exiting through the top of his head and completely transecting his frontal lobes. He lived for twelve years following the accident and his behavior was recorded in some detail, informing scientists about the unique function of the frontal lobes and their important role in personality and decision-making. The insights from such case studies have often generated hypotheses to be tested in subsequent manipulative experiments.

Fred H. Gage

Descriptive studies can also consist of the straightforward act of observing properties of the nervous system without manipulations. This type of research is usually the first crucial step in acquiring knowledge about a newly discovered gene, protein, neural subtype, or connection between neurons. Examples can be highlighted at every level of analysis. A novel gene can be sequenced and its expression pattern in the brain can be mapped, or the peptide sequence of a protein can be described and its distribution in the nervous system can be shown in great detail. In addition, a specific neuron can be described in terms of the genes and proteins it expresses, as well as its unique morphological characteristics and electrophysiological properties. On a broader scale, the connections between groups of neurons can be elucidated, describing both their input to their respective dendrites and spines and also their outputs, by way of the axons and boutons. Once the anatomical properties of a network are described, the electrochemical properties of their connections and network can be revealed.

These descriptive studies are excellent for generating hypotheses about the function of the brain at all levels of analysis. Once there is enough basic information to generate a coherent hypothesis about the function of some level of the brain – for example, an anatomical pathway in the brain that is responsible for our ability to recognize a face – we then want to test if the pathway is required for facial recog-

nition. In all areas of biological sciences and at all levels of analysis, testing a hypothesis is achieved through *gain-* and *loss-of-function* experiments. In a loss-of-function experiment, the experimenter silences, blocks, disrupts, or turns off specific components of a proposed pathway in an attempt to determine the required elements for appropriate function. In some cases, the loss-of-function technique may not be precise, so to further track down the requirement for a component in the functional path, a gain-of-function experiment can be conducted to replace each of the components of the pathway that were disrupted in the loss-of-function experiment. Loss-of-function experiments can be conducted at all levels of function: to test the importance of specific genes in cells within the inner ear for specific components of hearing; to test the roles of specific regions of the temporal lobe in learning; or even to test the importance of sleep in the consolidation of memory.

This volume of *Dædalus* dedicated to the brain and nervous system cannot cover all aspects of this very deep and broad field of study; but we are fortunate to have recruited an outstanding group of active scientists to help us examine select subdivisions in the field of neuroscience. These authors and scientists are not only major contributors to their particular areas of focus, but are experienced communicators with track records of explaining and translating complex concepts to intelligent readers and listeners outside of their discipline.

Robert Wurtz, in "Brain Mechanisms for Active Vision," presents a clear and lucid essay on the remarkable mechanisms behind our ability to see the world around us. In "Perceiving," Thomas Albright reveals how we change the sensory experience of vision into a cognitive perception and how this is regulated by other events in the environment. A. J. Hudspeth's essay, "The Energetic Ear," explains the dynamic inner workings of the ear and how sound waves are translated in the brain to allow us to hear. Larry Squire and John Wixted offer a primer on memory entitled "Remembering," based on both critical case studies and the experimental studies that have led to our current understanding. And in their essay "Sleep, Memory & Brain Rhythms," Brendon Watson and György Buzsáki provide a coherent and provocative study of the importance of sleep in our memory and how the rhythmic activity in the circuits of the brain may control the relationship between sleep and memory.

Emilio Bizzi and Robert Ajemian have contributed the essay "A Hard Scientific Quest: Understanding Voluntary Movements," which explains both the basics of how we move through our environment and how movement is regulated by sensory experience. "Feelings: What Are They & How Does the Brain Make Them?" – Joseph LeDoux's addition to the volume – describes the fundamentals of emotional behaviors both from the behavioral perspective and from the neurobiological basis of emotions. Earl Miller and Timothy Buschman, in their essay "Working Memory Capacity: Limits on the Bandwidth of Cognition," discuss cognitive capacity, with a special focus on processing limitations rooted in oscillatory brain rhythms ("brain waves"). Finally, in his essay "Consciousness," Terry Sejnowski tackles the slippery concept of consciousness and helps us understand the difference between being aware and being consciously aware.

Although this volume cannot extend to all sensory and motor systems and their integration, we are hopeful that this sampling of neuroscience will encourage you to read more on these exciting topics, and we hope we will be able to return to *Dædalus* with additional volumes on neuro-

science. More specific, we have not here considered what happens when the brain is damaged or aged, or when genetic errors occur. A volume on "The Brain and its Disorders" is currently in the planning stages at the American Academy; in the meantime, we hope this collection provides a foundation for you to learn about the brain and whets your appetite for more.

Fred H. Gage

Brain Mechanisms for Active Vision

Robert H. Wurtz

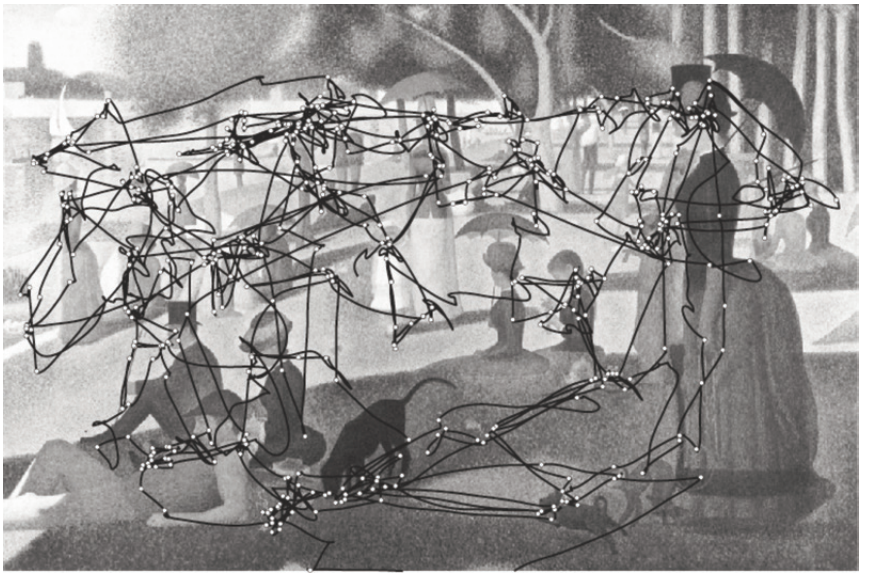
Abstract: Active vision refers to the exploration of the visual world with rapid eye movements, or saccades, guided by shifts of visual attention. Saccades perform the critical function of directing the high-resolution fovea of our eyes to any point in the visual field two to three times per second. However, the disadvantage of saccades is that each one disrupts vision, causing significant visual disturbance for which the brain must compensate. Exploring the interaction of vision and eye movements provides the opportunity to study the organization of one of the most complex, yet best-understood, brain systems. Outlining this exploration also illustrates some of the ways in which neuroscientists study neuronal systems in the brain and how they relate this brain activity to behavior. It shows the advantages and limitations of current approaches in systems neuroscience, as well as a glimpse of its potential future.

ROBERT H. WURTZ, a Fellow of the American Academy since 1990, is a National Institutes of Health Distinguished Investigator in the Laboratory of Sensorimotor Research at the National Eye Institute. He is a member of the National Academy of Sciences and former President of the Society for Neuroscience. His work on visual and oculomotor systems in primates and humans has been published in such journals as *The Journal of Neuroscience*, *Journal of Neurophysiology*, *Nature*, and *Science*.

Though our perception convinces us beyond a doubt that we see the visual world as one coherent whole, we actually see a series of snapshots from which we construct a unified view of the world in our brains. Figure 1, which shows a record of the eye movements of a viewer inspecting the Georges Seurat painting *A Sunday Afternoon on the Island of La Grande Jatte*, illustrates the snapshot process. A record of the viewer's eye movements is superimposed on the painting. The black lines show the path of the eyes as they move from one part of the painting to another. These rapid eye movements, referred to as saccades, are not only fast but frequent, occurring two to three times per second. The dots at the end of each saccade are visual fixations, the points at which the eyes come to rest. Nearly all of our useful vision occurs during fixations, because the scene is then stationary in front of the eyes. With successive fixations, the brain receives a series of snapshots of different fragments of the scene. From these fragments, we become convinced that we see the whole scene at once.

doi:10.1162/DAED_a_00314

Figure 1
Examination of a Scene with Eye Movements of the Viewer Superimposed



Saccades are represented by black lines, and intervening periods when the eyes are stationary (fixations) are represented by white circles. The human viewer documented here looked at this painting, *Sunday on the Island of the La Grande Jatte* by Georges Seurat, for a period of three minutes. Note that the fixations are not to random locations, but rather to points of likely interest. A print of the painting was projected on a screen in front of the subject. Source: Figure prepared by the author.

Why bother with all these saccades? Why not just hold the eye steady and inspect the painting? The reason lies in the resolving power of the retina, the receiving surface within our eyes. The retina is equipped with receptors that respond to light, dark, and color, but it does not have uniform resolution across its surface. The highest receptor density is found in the central region of the retina, called the fovea, which gives us the highest visual resolution and enables us to see small details. Retinal areas outside this central region, responding to light from the periphery, have a lower density of receptors and therefore lower resolution. Thus, the viewer enjoying the Seurat painting is essentially using the fovea to examine the rich detail, jumping from one part of the painting to another. In the three minutes of saccades shown in Figure 1, the viewer examines the details in a substantial fraction of the painting, but never sees the whole scene at once.

How do subjects pick the next object to examine? Using their peripheral retina, subjects can see objects in the field at a relatively low resolution and select ones of potential interest to examine next. This shift of attention from one item to the next accompanies each saccade. This selection is not random: if we look at the saccades superimposed on the Seurat painting (Figure 1), we can see that the trees at the top of the painting and the grass at the bottom are largely ignored; in contrast, faces, dresses, and other significant objects are frequently inspected. The Russian psychologist Alfred Yarbus was the first to note the connection between eye movements and

objects of potential interest to the subject.[1] He concluded that the location of saccades was a good indicator of the subject's attentional shifts. Thus, the lines superimposed over the Seurat painting not only show saccades, but shifts of attention as well. Whether a particular object attracts attention is determined by a combination of the salience of the object (its color, motion, shape, brightness) and the subject's goals at a given moment. Attention is a critical factor in active vision – some might say the *most* critical factor – because it determines where we look. Furthermore, the neuronal activities in the brain related to shifts in attention and the generation of saccades are closely intertwined.[2] The term *active vision*, therefore, denotes the active exploration of the visual world with rapid saccadic eye movements from one point to another, each guided by a shift of attention. The goal of these saccades is to bring images to the fovea for detailed analysis. Even though these saccades displace the image of the whole visual field on the retina, the system operates so perfectly that we regard the scene as serenely stable.

Active vision comprises several functions, but of course, it is only a small part of the larger brain systems involved with sensation and motor control.[3] In turn, the puzzle of how these systems operate to produce action is just one of many global questions about how the brain produces all behavior, including learning, memory, and emotion; and even how consciousness arises from brain activity. Considering active vision alone, however, exemplifies the classic approach of reducing the overwhelming complexities of the brain to more easily understood fragments. Galileo had to study the solar system before we could study the universe.

The rest of this essay illustrates sequential steps in the investigation of the brain functions that underlie active vision (as we have outlined it above). The first step is to move from describing the benefits of saccades to specifying the problems that these saccades produce for vision. Next, we consider the logic of a particular brain mechanism called *corollary discharge*, which might eliminate these problems. We then briefly review our basic knowledge of the brain pathways that process visual stimuli and produce saccadic eye movements in order to locate the source of a corollary discharge. Finally, we consider how the corollary discharge circuits we find act to minimize the disruptions generated by saccades. This is a progress report, evaluating what we know about active vision and what we do not.

Every saccade produces two major problems for vision: blurring of the image and displacement of the image. Blurring occurs when the image of the scene is swept quickly across the retina during each saccade. Image displacement is closely related to blurring: each time the saccade moves the fovea from one part of the visual scene to another, the image on the retina is displaced (imagine a movie camera moving in jerks from one spot to another). For example, suppose you are looking at the dog in the foreground of Figure 1 and make a saccade that moves the fovea to one of the umbrellas. Not only does the image on the fovea change, but everything on the retina is also displaced.

These problems should be devastating to our vision, but they are rarely reported. Some individuals are aware of the blur that occurs during saccades, but no one is aware of the displacement of images on the retina. If these displacements were visible, most people would be seasick from looking at the lurching scene in front of them. These problems are immense because they arise at the very source of all visual knowledge: the retina. The solutions to these problems are also among the brain's most remarkable feats of information processing.

Philosophers and scientists have speculated for centuries about why we experience visual stability rather than seasickness with rapid eye movements.[4] A solution that emerges from many of their writings is that there must be a mechanism in the brain that warns the sensory systems that a movement is about to occur. With this warning, the sensory systems would be informed of the impending disruption and could compensate for its consequences. Hermann von Helmholtz, one of the premier visual scientists of the nineteenth century, referred to this internal signal as an "effort of will," while scientists who subsequently studied the problem in the twentieth century referred to it as an "efference copy" or "corollary discharge."[5] While the labels all represent the same phenomenon, I prefer *corollary discharge* because it suggests internal information about movement at all levels of the brain, not just at the level of movement output.

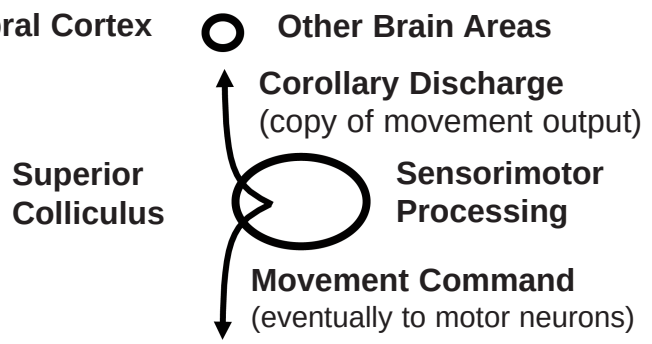
The logic of the corollary discharge is outlined in Figure 2. The basic principle is that the same sensorimotor processing area that drives the movement – in this case the saccadic eye movement – also produces a copy or corollary of this drive. The signals are identical, but one leads to the movement command, while the other is directed toward other brain areas to inform them that the eye is about to move. Recipient brain areas include those devoted to processing visual information, since they would receive the visual consequences of the saccade. For saccadic eye movements, both the movement command that produces the saccade and the corollary discharge circuits almost certainly originate in the superior colliculus (a subcortical area deep within the brain).[6] The corollary discharge is directed to the cerebral cortex, which covers the surface of the brain and is the site of the highest levels of information processing underlying much of our behavior. The concept of a corollary discharge is centuries old, but it is only in the last decade that we have begun to identify its neuronal implementation in brains similar to our own.

But what of the neuronal circuits that underlie the corollary discharge in the human brain? The usual assumption is that such circuits can be identified using brain imaging, particularly functional magnetic resonance imaging (fMRI). However, because fMRI signals are derived from changes in blood flow that are averaged over several seconds, fMRI is not adequate for investigating active vision. Over that period, the eye will have moved four to six times, changing the image on the retina with each movement and precluding fMRI records from capturing the rapid developments of active vision. But if imaging does not help, how do we begin to understand the brain activity underlying our active vision?

The answer is to study the organization of neurons in animals whose vision and eye movements are similar to ours. For active vision, the animal of choice has been the Old World monkey (the Rhesus monkey). These monkeys' visual discriminations and their range of eye movements are in most cases virtually identical to those of humans. The relevant anatomical connections in human and monkey brains are also remarkably similar. Techniques have been developed to painlessly record from brain neurons in monkeys while they perform a series of visual tasks that reveal changes in the brain during active vision. At this point, it is fair to say that most of what we know of the structure and function of the human brain for active vision is derived from studying the brains of Old World monkeys.

To determine where in the brain the corollary discharge signal originates, we first need to understand the basic organization of the pathways underlying active vision.

Figure 2
The Logic of the Corollary Discharge



A copy of the same signal sent to the eye movement motor neurons is sent to the cerebral cortex to inform it that the eyes are about to move. Source: Modified from R. H. Wurtz, "Neuronal Mechanisms of Visual Stability," *Vision Research* 48 (2008): 2070 – 2089.

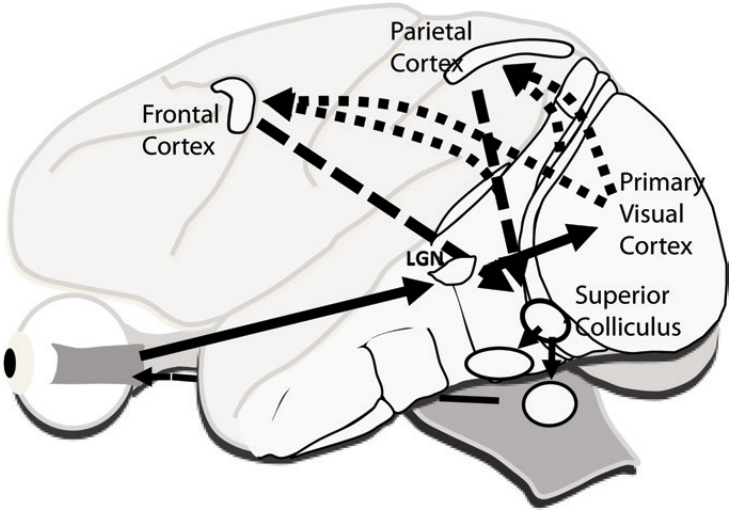
Experiments in dozens of laboratories around the world have built up an outline of the brain systems that underlie the analysis of visual input, the transition between visual and motor systems, and the motor output – all key phases of brain activity supporting active vision.[7]

Figure 3 outlines the major pathways in the monkey brain for active vision. Information from the eye passes through a relay nucleus and reaches the primary visual cortex (solid arrows). This visual area includes both the primary visual cortex and at least forty visual areas (not shown) that process different aspects of vision including shape, motion, depth, and color.[8] From these visual areas, projections go to the highest levels of the cerebral cortex – the parietal and frontal cortex (dashed arrows) – which are more directly related to the control of movement, including saccadic eye movements. Projections from these regions of the cortex then reach many subcortical structures (wide dashed arrows) that lie in the brainstem (roughly the region of the brain between the cortex and the spinal cord). The major target is the superior colliculus, the outputs of which eventually reach the eye motor neurons that activate the eye muscles (solid arrows again).

This outline from retinal input to eye movement output is a condensed version of the complete circuit that constitutes the neuronal basis of our visual perception and visual control of our eye movements. It is a working hypothesis that the neurons in the pathway in the monkey brain have activity identical to that in the human brain. But given that monkeys and humans have similar saccadic output for a given visual input, it is a reasonable hypothesis.

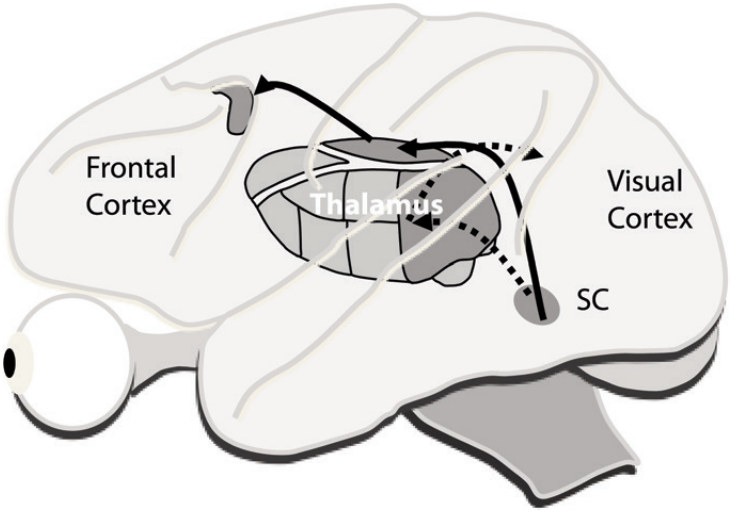
Where in the pathways underlying active vision might a corollary discharge arise? We are looking for an area with projections that convey a copy of the saccadic eye movement command to the cerebral cortex. The superior colliculus is such a site: it is the origin of commands to move the eye and a source of projections back to the cortex. Figure 4 shows the circuit for two corollary discharge pathways from the superior colliculus to the cerebral cortex. One pathway projects to the frontal cortex and originates in the layers of the superior colliculus where neurons are active before each saccade. A second pathway projects to the posterior visual regions of cortex and

Figure 3
Side View of the Monkey Brain Showing Circuits in the Brain for Visually Guided Saccades



The schematic outline of the pathway to the visual cortex is represented by solid arrows, the intervening pathways to the frontal and parietal cortex and then to the brainstem by dashed arrows, and the motor pathway to the eye muscles again by solid arrows. Source: Modified from R. H. Wurtz, "Neuronal Mechanisms of Visual Stability," *Vision Research* 48 (2008): 2070 – 2089.

Figure 4
Two Corollary Discharge Circuits to the Cerebral Cortex



The corollary discharge to the frontal cortex comes from saccade-related neurons in the superior colliculus (SC) and passes through the thalamus. The corollary discharge to the visual cortex is from visual neurons in the superior colliculus (SC) and also passes through the thalamus. Source: Modified from R. H. Wurtz, K. McAlonan, J. Cavanaugh, and R. A. Berman, "Thalamic Pathways for Active Vision," *Trends in Cognitive Science* 15 (2011): 177 – 184.

originates from different layers of the superior colliculus in which neurons respond to visual stimuli. As we shall see, these two pathways contribute to solving the problems generated by saccades: the first compensates for displacement on the retina; the second suppresses blur during saccades.

These projections to the cortex, whether from outside the brain (via sensory pathways) or inside the brain (via corollary discharge pathways), have a relay in the largest group of nuclei in the brain, the thalamus. Each of the senses and each internal signal have dedicated nuclei in the thalamus; thus, it is a basic feature of primate brain organization. To use the terms of air travel, every passenger flying to the cortex must change planes at the thalamus.

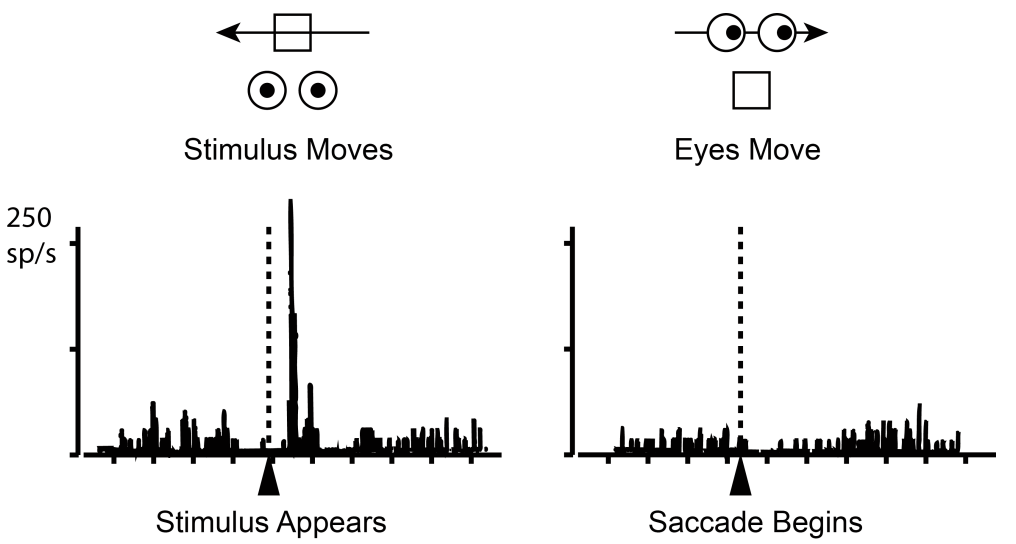
Having identified two corollary discharge pathways to the cortex, we can now investigate how they contribute to solving the two problems that saccades create for vision: blur and image displacement. The first problem is the blur produced by each saccade. The solution is relatively simple: suppress visual activity during the saccade and blur is suppressed as well. Early in the twentieth century, such suppression was thought to be produced by a "central anesthesia" in which activity during the saccade was simply blanked out.[9] When it became possible in the 1960s to record directly from monkeys' brains, it immediately became clear that neurons in the brain continue to respond to visual input quite well during saccades. There is, however, some suppression of neuronal activity during saccades, and as is the case for many biological problems, the brain provides more than one way of producing this suppression.

One mechanism that reduces blur is a purely visual phenomenon called visual masking: when a dim object is preceded or followed by a bright object, the dim object is not seen. Similar masking effects can be seen acting on neurons in the primary visual cortex.[10] This occurs only in a well-lit environment, though it is highly effective. It is not specifically related to eye movements, however, because masking of dim objects can occur any time bright objects are also present, even in the total absence of eye movements.

Corollary discharge also contributes to the suppression of neuronal activity in neurons, as illustrated for a superior colliculus neuron in Figure 5. The neuron responds to a stimulus moving in front of the stationary eyes (left) but not when the moving eyes pass over a stationary stimulus (right). The response is only suppressed during the eye movement and the corollary discharge is only present when the eye moves. The suppression therefore is correlated with the saccadic eye movement.[11] This suppression must be driven by a corollary discharge rather than visual masking, because masking only functions in the light and the corollary discharge–related suppression was demonstrated in the dark. Further experiments indicate that this suppression in the superior colliculus is passed through the thalamus to the visual cortex (Figure 4). The suppression can therefore act on the visual processing in the visual cortex that is thought to underlie visual perception.

The second problem for vision that may be ameliorated by the action of a corollary discharge is the displacement of images on the retina with each saccade. But the neuronal mechanisms are substantially more complicated than those used for blur suppression. To understand these mechanisms that may deal with image displacement, we must briefly consider how visual information is organized in the brain. Visual processing starts when the image of a visual scene falls on the retina and forms a retinotopic map; that is, a layout of the visual scene is mapped on the retina just as it exists in reality. As we have already

Figure 5
Neuronal Activity Suppressed During Saccades



Left: A superior colliculus neuron responds to a stimulus moved at eye-movement speed across its stationary receptive field. The response of the neuron on multiple stimulus sweeps is shown as a histogram of neuron responses (vertical axis) against time (horizontal axis). Right: For the same neuron, the same stimulus is now stationary. The saccadic eye movements sweep the receptive field across the stimulus, but there is no response. A corollary discharge acts on the neuron to suppress its activity during a saccade, thereby reducing the blurred stimulus that would otherwise be seen. Source: Modified from D. L. Robinson and R. H. Wurtz, "Use of an Extraretinal Signal by Monkey Superior Colliculus Neurons to Distinguish Real from Self-Induced Stimulus Movement," *Journal of Neurophysiology* 39 (1976): 852 – 870.

noted, the problem is that the whole map moves with each saccade and the brain must deal with the disruption in order to produce a stable visual perception. The problem could be solved in at least two ways.

One solution is for the brain to convert the retinotopic map into a spatial map higher in the chain of visual processing and continuously update it. After each saccade, the part of the visual scene around the fovea would be transferred to the spatiotopic map. The central part of the current scene would simply be pasted into the map at a location determined by the corollary discharge of the last saccade. This spatial map would be our brain's reconstruction of the visual world and would be used for all subsequent visual processing of a scene. If our brains employed such a spatial map, the displacement problem would automatically be solved, because no displacement ever occurs on this higher-order spatial map. The fatal problem with this theory, however, is that no evidence for such a map has been identified in the last forty years of research on vision in monkeys. Researchers *have* found some hints of conversion to spatial coordinates, and the map itself may have been so elusive because of the way it is represented in the brain. But at this time, we lack convincing evidence that a spatiotopic map exists in the brain, despite the simple elegance of this conceptual solution for visual stability.

Another explanation of perceived visual stability is that the retinotopic map is simply updated after each saccade. This idea

Brain Mechanisms for Active Vision

is based on experimental observations of single neurons in the parietal and later the frontal cortex.[12] In these experiments, neurons in these regions were found to have limited visual receptive fields (the areas in the visual field where visual stimuli activate neurons). As the monkey fixated on an object, a light was flashed in the receptive field. As expected, this produced a visual response (Figure 6, left). However, as the monkey prepared to make a saccade, a stimulus was flashed in the location that the receptive field would occupy after the saccade. Unexpectedly, this also activated the neuron (Figure 6, right). Why should activity be evoked from the site of the future receptive field even before the saccade was made? One possibility is that this future-field activity provides anticipatory information about an impending saccade and the location of the receptive field. Each of these pieces of information could be provided by a corollary discharge. With this anticipatory signal, the monkey would be forewarned that the impending visual disruption was due to its own saccade and not to something that happened in the outside world. We now know that this anticipatory activity in the frontal cortex is in fact dependent upon the corollary discharge, because if the corollary discharge is perturbed, the anticipatory activity of the frontal cortex neurons is greatly reduced.

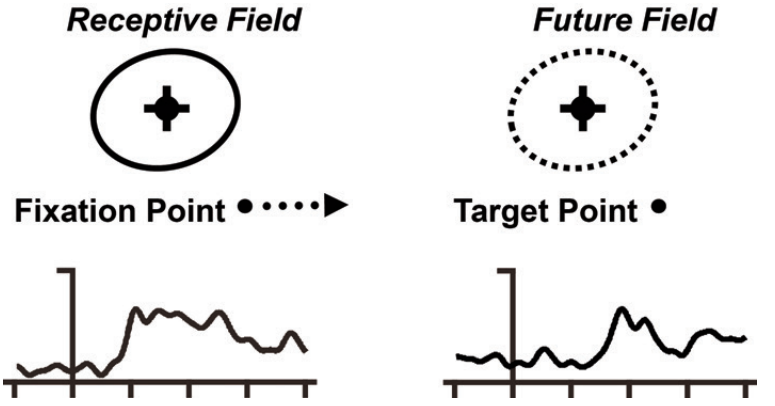
The brain mechanisms by which this anticipation might work to produce visual stability are not known. One hypothesis suggests that a match occurs between the anticipatory activity in the future receptive field before the saccade and the receptive field after the saccade (this match can only occur when a saccade is made).[13] While this hypothesis has not been tested experimentally, the necessary components of the hypothesis – a corollary discharge and future field activity – are well established. Clearly, the analysis of neuronal mechanisms is still at an early stage.

In conclusion, active vision is one of the first examples of a system in our brain that is neither sensory nor motor; rather, its whole function depends on the synchrony between visual input, movement output, and systems internal to the brain, such as those for corollary discharge. This organization allows primates to use the high-resolution fovea to examine anything anywhere in the visual field. The price paid is that eye movements generate major problems for the visual system two to three times per second. While these problems originate at the very first stages of the visual system, the solutions do not: they are spread throughout the visual system across wide regions of the brain, including the highest levels of visual processing in the parietal and frontal cortex.

Active vision creates two significant problems – blurring and displacement of retinal images – and there are two identified brain circuits that might contribute to the solutions. Both circuits rely on a corollary discharge derived from the brain areas where commands to move the eyes originate. One circuit extends from the superior colliculus to the visual cortex and may provide an input to the cortex that suppresses blur during saccades. The second circuit connects the superior colliculus and the frontal cortex and may provide the anticipatory activity that warns the frontal cortex that a disruption of visual input is about to occur as a result of the subject's own eye movements. While we have identified two problems and two circuits, it would be premature to assume that either problem is solved exclusively by these circuits without exploring other possibilities. The work so far should be viewed as an enthusiastic start, not an assured solution.

So far, a corollary discharge has been identified in other branches of visual-motor systems, but it has not been identified in other systems in the primate brain. There

Figure 6
Anticipatory Activity of a Neuron before a Saccade that Might be Related to the Compensation for Image Displacement



Left: Response of a frontal cortex neuron to stimuli flashed in its receptive field during fixation. The smoothed histogram shows the neuron's response. Right: Response to stimuli flashed at the receptive field's future location. The flashes occur before the saccade; thus, the responses are in anticipation of it. Such an anticipatory signal informs the brain that the coming visual displacement is the result of the subject's own saccade, not of some motion in the outside world. Source: Modified from M. A. Sommer and R. H. Wurtz, "Influence of the Thalamus on Spatial Visual Processing in Frontal Cortex," *Nature* 444 (2006): 374 – 377.

is every reason to believe, however, that a corollary discharge will be found in other systems and at multiple levels within those systems. This expectation is strengthened by the extent to which corollary discharge is used throughout the animal kingdom.[14]

Finally, the process used to explore the neuronal basis of active vision is an excellent illustration of how neuroscientists go about identifying the neuronal mechanisms that underlie behavior. As we have seen, the study begins with the quantification of behavior that is straightforward for active vision, because methods for measuring eye movements and psychophysical measurements of visual perception are both readily available. The second step is to establish a correlation between the measured behavior and neuronal activity in the brain. Analysis of active vision, for example, depends on the extensive correlations established between behavior and neuronal activity undertaken over the last century, first on the visual system and then, when recording eye movements became possible, on the visual-motor system. From what we know now, the example of active vision demonstrates unequivocally that future progress depends first and foremost on knowing the basic organization of the relevant brain circuits. Without that knowledge, treatment of vision-related human disease is likely to proceed at a glacial pace or simply be fruitless.

After an understanding of basic brain systems has been developed, a third step then becomes possible: extracting and identifying a specific circuit – such as that for a corollary discharge – from the massive number of connections within the brain. Once the specific circuit has been identified, neuronal activity in the circuit can be interrupted in order to see how behavior is changed, which allows more specific evaluation of the circuit's function. The final step, and frequently the most elusive one, is to develop a precise model that represents the elements of the system

and that predicts its functions: both those known and those as yet unrecognized.

We currently have only a glimpse of the neuronal basis of active vision in the brain. However, even with the limited methods we have now, an understanding of the system's organization seems reachable. The hope is that active vision becomes an example of how a complex problem is solved by the brain in simple and clever – but not necessarily intuitive – ways.

ENDNOTES

1 A. L. Yarbus, *Eye Movements and Vision* (New York: Plenum, 1967).

2 Michael E. Goldberg and Robert H. Wurtz, "Activity of Superior Colliculus in Behaving Monkey II. Effect of Attention on Neuronal Responses," *Journal of Neurophysiology* 35 (4) (1972): 560 – 574; and Tirin Moore, Katherine M. Armstrong, and Mazyar Fallah, "Visuomotor Origins of Covert Spatial Attention," *Neuron* 40 (4) (2003): 671 – 683.

3 Robert H. Wurtz, "Neuronal Mechanisms of Visual Stability," *Vision Research* 48 (20) (2008): 2070 – 2089.

4 Otto-Joachim Grösser, "On the History of the Ideas of Efference Copy and Reafference," *Clio Medica* 33 (1995): 35 – 55.

5 R. W. Sperry, "Neural Basis of the Spontaneous Optokinetic Response Produced by Visual Inversion," *Journal of Comparative Physiology and Psychology* 43 (6) (1950): 482 – 489; Hermann von Helmholtz, *Helmholtz's Treatise on Physiological Optics*, 3rd ed., trans. James Powell Cocke Southall (New York: Optical Society of America, 1910); and E. von Holst and H. Mittelstaedt, "Das Reafferenzprinzip: Wechselwirkungen zwischen Zentralnervensystem und Peripherie," *Naturwissenschaften* 37 (1950): 464 – 476.

6 Mark A. Sommer and Robert H. Wurtz, "Brain Circuits for the Internal Monitoring of Movements," *Annual Review of Neuroscience* 31 (2008): 317 – 338.

7 For summaries of the visual pathways and the oculomotor pathways, see Eric R. Kandel, James H. Schwartz, Thomas M. Jessell, Steven A. Siegelbaum, and A. J. Hudspeth, *Principles of Neural Science*, 5th ed. (New York: McGraw-Hill, 2012), ch. 27 – 29 and ch. 39, respectively.

8 For a discussion of the areas of the visual cortex in monkeys and humans, see Daniel J. Felleman and David C. Van Essen, "Distributed Hierarchical Processing in the Primate Cerebral Cortex," *Cerebral Cortex* 1 (1) (1991): 1 – 47; and David C. Van Essen, James W. Lewis, Heather A. Drury, Nouchine Hadjikhani, Roger B.H. Tootell, Muge Bakircioglu, and Michael I. Miller, "Mapping Visual Cortex in Monkeys and Humans Using Surface-Based Atlases," *Vision Research* 41 (10 – 11) (2001): 1359 – 1378.

9 Frances C. Volkmann, "Human Visual Suppression," *Vision Research* 26 (9) (1986): 1401 – 1416.

10 For examples of masking in the primary visual cortex, see Stuart J. Judge, Robert H. Wurtz, and Barry J. Richmond, "Vision During Saccadic Eye Movements. I. Visual Interactions in Striate Cortex," *Journal of Neurophysiology* 43 (1980): 1133 – 1155; and Stephen L. Macknik, "Visual Masking Approaches to Visual Awareness," in *Visual Perception, Part 2: Fundamentals of Awareness: Multi-Sensory Integration and High-Order Perception*, ed. Susana Martinez-Conde, S. L. Macknick, Maria M. Martinez, Jose-Manuel Alonso, and Peter U. Tse (Amsterdam: Elsevier Science B.V., 2006), 177 – 215.

11 For examples of neuronal suppression with saccades, see David L. Robinson and Robert H. Wurtz, "Use of an Extraretinal Signal by Monkey Superior Colliculus Neurons to Distinguish Real from Self-Induced Stimulus Movement," *Journal of Neurophysiology* 39 (1976): 852 – 870; and Rebecca A. Berman and Robert H. Wurtz, "Signals Conveyed in the Pulvinar Pathway from Superior Colliculus to Cortical Area MT," *The Journal of Neuroscience* 31 (2011): 373 – 384.

[12] For examples of anticipatory neuronal activity before saccades, see Jean-René Duhamel, Carol L. Colby, and Michael E. Goldberg, "The Updating of the Representation of Visual Space in Parietal Cortex by Intended Eye Movements," *Science* 255 (5040) (1992): 90 – 92; Marc A. Sommer and Robert H. Wurtz, "Influence of the Thalamus on Spatial Visual Processing in Frontal Cortex," *Nature* 444 (7117) (2006): 374 – 377; and Marc M. Umeno and Michael E. Goldberg, "Spatial Processing in the Monkey Frontal Eye Field. I. Predictive Visual Responses," *Journal of Neurophysiology* 78 (3) (1997): 1373 – 1383.

[13] Heiner Deubel, Werner X. Schneider, and Bruce Bridgeman, "Transsaccadic Memory of Position and Form," *Progress in Brain Research* 140 (2002): 165 – 180.

[14] Trinity B. Crapse and Marc A. Sommer, "Corollary Discharge across the Animal Kingdom," *Nature Reviews Neuroscience* 9 (2008): 587 – 600.

Perceiving

Thomas D. Albright

Abstract: Perceiving is the process by which evanescent sensations are linked to environmental cause and made enduring and coherent through the assignment of meaning, utility, and value. Fundamental to this process is the establishment of associations over space and time between sensory events and other sources of information. These associations provide the context needed to resolve the inherent ambiguity of sensations. Recent studies have explored the neuronal bases of contextual influences on perception. These studies have revealed systems in the brain through which context converts neuronal codes for sensory events into neuronal representations that underlie perceptual experience. This work sheds light on the cellular processes by which associations are learned and how memory retrieval impacts the processing of sensory information. Collectively, these findings suggest that perception is the consequence of a critical neuronal computation in which contextual information is used to transform incoming signals from a sensory-based to a scene-based representation.

THOMAS D. ALBRIGHT, a Fellow of the American Academy since 2003, is Professor and Director of the Vision Center Laboratory and the Conrad T. Prebys Chair in Vision Research at the Salk Institute for Biological Studies. His work has appeared in such journals as *Nature*, *Science*, *Neuron*, *Journal of Neurophysiology*, and *The Journal of Neuroscience*.

Perceiving is a common English word with a number of related colloquial meanings: it is the act of understanding, realizing, seeing, noticing, or becoming aware of. In modern neuroscience, our working definition of perception is captured well by the *Oxford English Dictionary*: "The action of the mind by which it refers its sensations to an external object as their cause." This definition has roots in the corpus of eighteenth-century philosophy – beginning with George Berkeley and David Hume – was expanded upon by later British associationists, and became a foundation of both experimental psychology and modern neuroscience.

There are two essential features of this definition, the first being the distinction between *perception* and *sensation*. Sensation is the immediate neurobiological consequence of stimulating *sensory transducers*[1] such as photoreceptors, mechanoreceptors, and chemoreceptors. Sensory events are ubiquitous and can affect behavior directly – the spinal reflex of pulling your hand back from a hot surface is one simple example – but they are fleeting, discontinuous, and lacking semantic content. Perception enriches sensation by reference to other knowledge or experi-

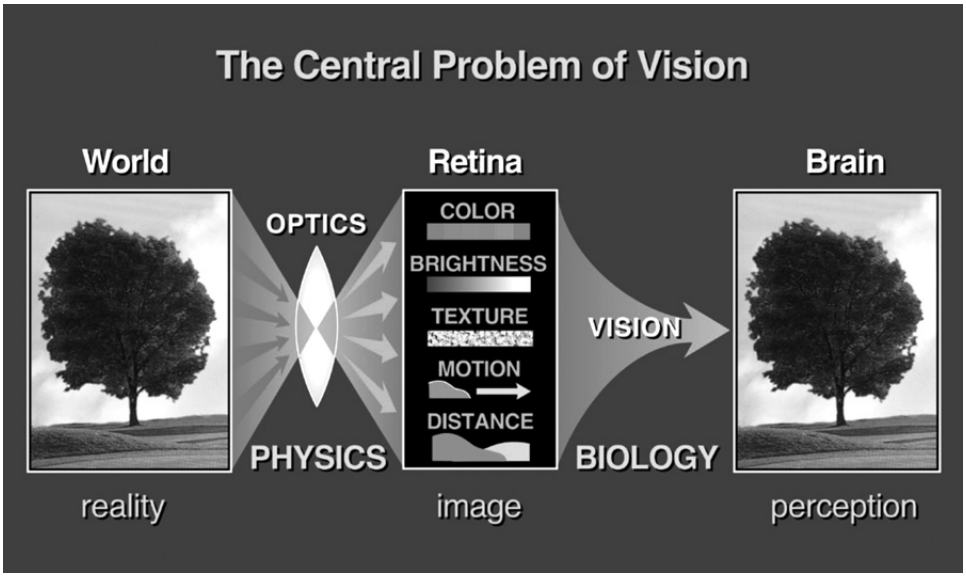
doi:10.1162/DAED_a_00315

ence. In the words of British associationist John Stuart Mill, "perception reflects the permanent possibilities of sensation"[2] – things as they were or might be – and in doing so reclaims from evanescent sensory events the enduring structural and relational properties of the world. The philosopher and psychologist William James expressed a similar view in his discussion of "The Perception of Things": "Perception thus differs from sensation by the consciousness of farther facts associated with the object of the sensation."[3] Or in the prescient words of nineteenth-century perceptual psychologist James Sully: the mind "supplements a sense impression by an accompaniment or escort of revived sensations, the whole aggregate of actual and revived sensations being solidified or 'integrated' into the form of a percept."[4] Building on the philosophical traditions from which the discipline of psychology was born, William James further stressed the need for associative neuronal processes to achieve this integration: "The chief cerebral conditions of perception are the paths of association irradiating from the sense-impression."[5]

The second essential feature of neuroscientists' definition of perception concerns the attributes of the thing (or things) to which sensation is referred. In particular, the referent is viewed as the "cause" of sensation. This concept was developed by the eighteenth-century Scottish philosopher Thomas Reid, who argued that sensation "suggests to us" an object as its source: "We all know that a certain kind of sound suggests immediately to the mind *a coach passing in the street*."[6] As James later noted, perception is "the consciousness of particular material things present to sense."[7] This attribution to the source of sensation is profoundly important and meaningful for perception and behavior. To fully appreciate this importance, it is useful to consider the larger function of sensory systems and the computational problems faced by them. For a variety of reasons – some technical, some conceptual, and some historical – this consideration is easiest to undertake for the visual modality, and that is the approach I will use herein. The principles of perceiving and the underlying neuronal mechanisms revealed by the study of vision nonetheless have broad relevance to other sensory modalities.

Figure 1 illustrates what I call the "central problem of vision." Technically this is two problems, one of which is optical and the other biological. The optical problem involves reflection of light off of environmental surfaces, refraction of that light by the crystalline lens of the eye, and, finally, projection of the light onto the back surface of the eye – the surface known as the *retina*,[8] which is lined with neuronal tissue responsible for phototransduction – to form the retinal image. This image is merely a pattern of light that changes in intensity and wavelength over space and time. But our survival depends on our ability to engage with – to recognize and to navigate amongst – the features of the visual scene that gave rise to the retinal image. The biological problem, which is by far the more difficult of the two, thus involves reconstruction of the properties of the visual scene, given only the pattern of light present in the retinal image. This inverse problem of optics is an example of what is known formally as an ill-posed problem: it is a problem without a unique solution. Because of the dimensionality reduction that accompanies optical projection of the world onto the retina, there is simply not sufficient information present in the retinal image to uniquely identify its environmental causes. To put it bluntly, the number of visual scenes that could give rise to any specific retinal image is infinite.

Figure 1
The Central Problem of Vision



The problem is twofold: one component is optical and the other biological. The optical problem involves reflection of light off of surfaces in the visual environment. This light is refracted by the crystalline lens at the front of the eye, resulting in a pattern of light (the retinal image) that is projected and focused on the retinal surface at the back of the eye. The biological problem – the problem of *perception* – involves identification of the elements of the visual scene that gave rise to the retinal image. This is a classic inverse problem for which there is no unique solution: a given retinal image could be caused by any one of an infinite set of visual scenes.

This fundamental ambiguity is not reflected in human perceptual experience, however, for we generally arrive quickly at a solution and the chosen solution is nearly always correct (in that it reflects the "true" environment). This is only possible by using other sources of information, including the spatial relationships between different features of the image (spatial context), the observer's prior experiences (temporal context), and consistent properties of the visual world (for example, that light comes from above) that have become embedded in the computational machinery of the brain through natural selection (evolutionary context). Recent experiments summarized herein have revealed much about how these contextual influences enable perception. Source: Figure prepared by author.

In view of this intrinsic ambiguity, perhaps the most astonishing thing about visual perceptual experience is that we rarely have difficulty arriving at a unique solution for the environmental cause of the retinal image. Moreover, the solution we adopt is nearly always the correct one (at least within margins of error allowable for our behavioral interactions with the world). Those cases where we arrive at the wrong solution are what we call "illusions." Vision is able to accomplish reliable disambiguation of the retinal image by virtue of context, which, broadly speaking, consists of other pieces of information that are either 1) co-present in the retinal image (spatial context); 2) learned based on the observer's prior experience with the world (temporal context); or 3) embedded in the computational machinery of the brain as a result of evolution in an environment that has consistent and well-defined properties (evolutionary context). Context is thus James's "farther facts

associated with the object of sensation." Using the available context as clues, the process of disambiguation – the process of *perceiving* – is best characterized as probabilistic inference about the cause of sensation. Again according to James, "perception is of probable things."[9] Or to use a familiar colloquialism, "we generally see what we expect to see."

Context plays an extended role in perception in that it also helps resolve symbolic form. Sensory inputs are replete with symbolism – often quite abstract and multifaceted – and to perceive is to grasp the meanings of the symbols. Perhaps the most dramatic example of this is human language, which is by definition an experience-dependent mapping of auditory and visual stimuli onto meaningful objects, actions, and concepts. To illustrate how context resolves this mapping and, by doing so, makes perception possible, William James offered the phrase *Pas de lieu Rhone que nous*.[10] If the listener assumes that the spoken phrase is French, it is unintelligible. If, however, the listener is informed that the spoken phrase is English, the very same sounds are perceived as *paddle your own canoe*. James further noted that "as we seize the English meaning the sound itself appears to change." In other words, the percept is both reconciled and subjectively qualified by the context provided.

With these phenomenological aspects of perception in mind, research in recent years has focused on the mechanisms by which contextual cues modulate neuronal signals originating from the senses, thereby reflecting the thing perceived. In the case of vision, the question is: how do neuronal signals that initially reflect the properties of the retinal image become transformed (via context) such that they reflect instead the properties of the visual scene? Based on the discoveries presented below, I will argue that this transformation from an image-based representation to a scene-based representation is among the most basic and fundamental operations of the *cerebral cortex*[11] and is the neurobiological explanation for the fact that "perception reflects the permanent possibilities of sensation."

To provide a foundation for understanding the neuronal bases of perceiving, I will briefly review our present understanding of the organization of the primate visual system. Much of this information has come from the use of three key experimental approaches:

1) Neuroanatomy: These studies, which are made possible by methods for selective visual labeling of brain tissue, provide a picture of *neurons*[12] – the basic cellular elements of the brain – as well as a picture of the wiring diagram of neuronal connectivity.

2) Neurophysiology: These studies reveal the types of sensory signals carried by neurons and the manner in which those signals are transformed at each computational stage in the signal processing hierarchy. In practice, this information is recorded physiologically by evaluating neuronal *receptive field*[13] properties. The receptive field (RF) is a central concept in sensory neurobiology and is defined as the region of sensory space that, when stimulated, elicits a change in the activity of the recorded neuron (typically quantified as frequency of *action potentials*[14]). Visual RF properties may include, in addition to spatial location, sensitivity to complex spatial, temporal, and/or chromatic properties of light. A more recent concept important for this discussion of perception is that of the RF surround (sometimes known as the "non-classical RF"). A sensory stimulus that falls in a neuron's RF surround does not directly cause a change in activity (by definition), but rather has the ability to modulate activity driven by a stimulus in the RF. As we shall see, the

RF surround provides a mechanism by which contextual cues may influence (and disambiguate) neuronal responses to sensory stimuli.

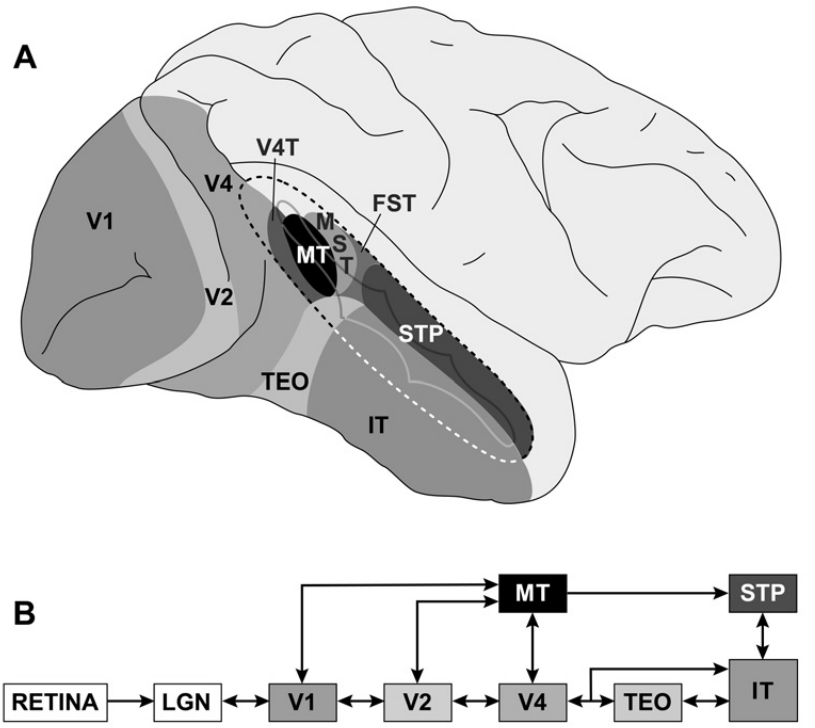
3) Behavioral analysis: This approach involves acquiring behavioral reports of subjective states, such as those associated with perceiving. These behavioral measures reveal correlations (and implied causal relationships) between a perceptual state and the neuronal signals recorded from specific anatomical circuit locations.

The first act of visual processing, which occurs in the retina, is phototransduction (the conversion of energy in the form of light into energy in the form of neuronal activity). Signals leaving the retina are carried by the *optic nerve/tract*[15] and convey information about the location of a stimulus in visual space, the wavelength composition of the stimulus, and about spatial and temporal contrast between different stimuli. Optic tract fibers *synapse*[16] in a region of the thalamus (a form of sensory transponder located in the center of the *forebrain*[17]), known as the *lateral geniculate nucleus*,[18] and the signals are passed on from there via a large fiber bundle to terminate in the *primary visual cortex*,[19] or *area V1* (Figure 2). Area V1 is located in the occipital lobe (the most posterior region) of the cerebral cortex, which is the massive convoluted sheet of neuronal tissue that forms the exterior of the forebrain. Signals recorded physiologically reveal that V1 neurons extract a number of basic features of the pattern of light in the retinal image, including position in visual space, color, contour orientation, motion direction, and distance from the observer. These signals ascend further to contribute to functionally specific processing – object recognition, spatial understanding, visual-motor control – in a number of additional visual areas, which collectively make up approximately one-third of the human cerebral cortex.

Attempts to understand how visual cortical neurons account for perceptual experience have been largely reductionist in approach. These experiments, which spearheaded the field of sensory neurobiology fifty years ago, typically involve manipulation of a very simple stimulus along a single sensory dimension – a contour is varied in orientation, a patch of light is varied in color, a moving texture is varied in directionality – and placed within the RF of a physiologically recorded neuron. This approach has revealed fundamental features of the ways in which basic attributes of a sensory stimulus are detected and encoded by neuronal activity. But in the end this approach has told us little about perception because the stimuli are devoid of meaning. They lack the context needed to identify environmental cause and are thus ambiguous in the most complete sense possible.

The alternative experimental approach is one in which the context is varied, such that the percept (the inferred environmental cause) is independent of the parameters of the sensory stimulus. The real world presents a rich set of such conditions, of course, but there is an advantage to an intermediate experimental approach in which perceptual and *neuronal responses*[20] to a simple well-defined stimulus – the "target" – are evaluated in the presence of simple contextual manipulations.[21] In our laboratory we have used both spatial and temporal context for this purpose. Spatial context is defined here as other (nontarget) features of the retinal image, such as the color, pattern, or motion of spatial regions surrounding the target (the "surround"). There are many well-known perceptual phenomena in which experience of a simple target stimulus is markedly influenced by the surround. Perception of brightness and color hue, for example, are heavily context dependent (Figure 3), often in complex and revealing ways. We have also discovered and explored a number of

Figure 2
Locations and Connectivity of Cerebral Cortical Areas of Rhesus Monkey (Macaca mulatta) Involved in Visual Perception



Because of the similarity of its visual functions to those of humans, the vast knowledge of the organization of its cerebral cortex, and the richness of its behavioral repertoire, the rhesus monkey has been an extremely powerful model for understanding the brain bases of visual perception.

(A) Lateral view of cerebral cortex of rhesus monkey. Front of brain is at right and top of brain is at top of image. Superior temporal sulcus is partially unfolded (dashed line) to show visual cortical areas that lie within. Differently shaded regions identify a subset (visual areas V1, V2, V4, V4T, MT, MST, FST, STP, TEO, IT) of the nearly three dozen distinct cortical areas involved in the processing of visual information.

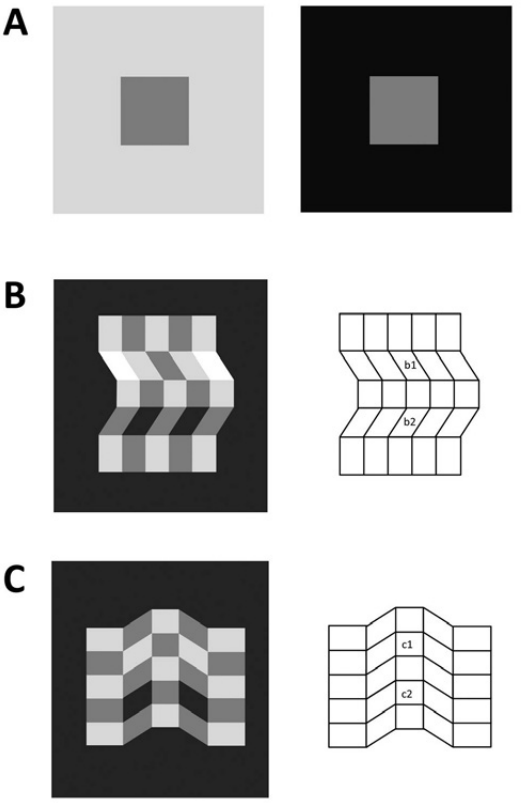
(B) Connectivity diagram illustrating a subset of known anatomical projections from retina to primary visual cortex (V1) and up through the inferior temporal (IT) cortex. Most projections are bi-directional.

Source: Figure prepared by author.

compelling effects of context on perceived motion.[22]

In all of these cases, the percept varies because the spatial context leads to different plausible interpretations of the cause of the sensory stimulus. My work with Gene Stoner (Salk Institute) on context-dependent visual motion perception illustrates this well.[23] Stoner designed a stimulus in which a diamond-shaped pattern of vertical stripes was viewed in the context of a textured surround (Figure 4). (We call this the "barber diamond" stimulus because it mimics features of the classic "barber pole" illusion, in which rotating stripes appear to move downward along the axis of the pole.) Portions of the textured surround were manipulated (using

Figure 3
Complex Influences of Spatial Context on Brightness Perception



(A) Simultaneous contrast enhancement. The intensities of the two central squares are identical, but the central square on the left appears darker. This brightness illusion, which has been known for centuries, can be accounted for by the differences in spatial context (light on left and dark on right).

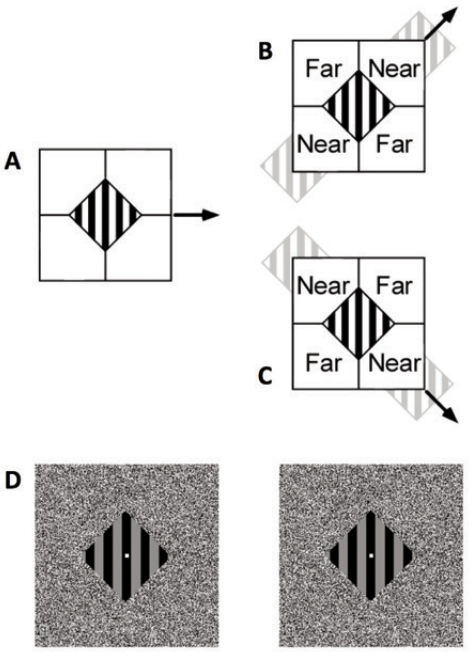
(B) The "corrugated plaid" illusion devised by vision scientist Edward Adelson shows the pronounced dependence of brightness perception on image cues for depth, form, and shading. Patches b1 and b2 are identical in intensity, but b2 appears brighter.

(C) By changing cues for depth, form, and shading, yet leaving the local luminance configuration intact, the brightness illusion present in (B) is greatly reduced. Patches c1 and c2 are identical in intensity and they appear approximately so.

The illusion can be accounted for by the use of contextual cues to identify the visual scene that produced the image. One important visual scene property is the reflectance of surfaces, since that property enables object recognition. The brightness values perceived in this illusion correspond to probabilistic inferences about surface reflectance that are driven by contextual cues that suggest a specific visual scene interpretation.

Source: Edward H. Adelson, "Perceptual Organization and the Judgment of Brightness," *Science* 262 (1993): 2042 – 2044 ; and Thomas D. Albright, "Why Do Things Look as They Do?" *Trends in Neuroscience* 17 (1994): 175 – 177.

Figure 4
Schematic Depiction of "Barber Diamond" Stimuli Used to Study the Influence of Context on Perceived Motion and Its Neuronal Correlates



(A) Stimuli consisted of a moving pattern of vertical stripes framed by a static, diamond-shaped aperture. The striped pattern itself was placed in the plane of ocular fixation while four textured panels that defined the aperture were independently positioned in depth using binocular disparity cues. The striped pattern was moved either leftward or rightward on each trial. Each of the four stimulus conditions used was a unique conjunction of direction of motion (right versus left motion of the striped pattern) and depth-ordering configuration. The two conditions illustrated (B and C) were created using a rightward moving pattern of stripes; two additional conditions (not shown) were created using leftward moving stripes. *Near* and *Far* identify the depth-ordering of the textured panels relative to the depth plane of the striped pattern.

(B) Upper-right and lower-left panels were placed in the *Near* depth plane, while the upper-left and lower-right panels were in the *Far* depth plane. Line terminators formed at the boundary of *Near* surfaces and the striped pattern are classified as extrinsic features resulting from occlusion, and the striped pattern is perceived to extend behind the *Near* surface. (Note that the gray stripes in this illustration are not part of the stimulus and are used solely to illustrate perceptual completion.) Conversely, line terminators formed at the boundary of the *Far* surfaces and the striped pattern are classified as intrinsic: they appear to result from the physical termination of the surface upon which the stripes are "painted." As a result of this depth-ordering manipulation and ensuing feature interpretation, observers typically perceive the moving pattern of stripes in (B) as belonging to a surface that slides behind the *Near* panels and across the *Far* panels (to the upper-right). This direction is identified with motions of intrinsic terminators.

(C) The condition shown contains the same rightward motion of the pattern of stripes, but employs the depth-ordering configuration that is complementary to (B). In this case, observers typically perceive the striped pattern as belonging to a surface that is drifting to the lower right.

(D) 3-D illustration of one depth-ordering configuration used for these experiments. The 3-D percept requires "free cross-fusing" of the left and right panels (such that the right eye is aimed at the left panel and the left aimed at the right). Those viewers capable of free fusing should perceive the 3-D layout of the stimulus. In practice, the central striped pattern also drifted leftward or rightward.

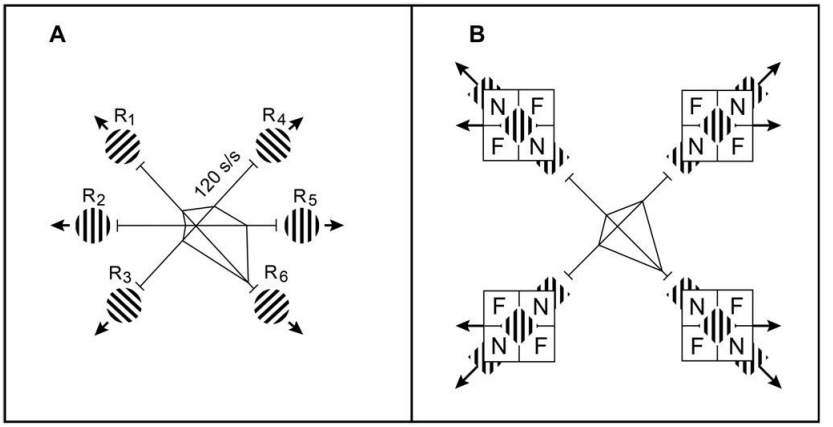
A demonstration of this depiction, with motion, is available at http://vcl-s.salk.edu/Research/Motion-Integration/. Source: Robert O. Duncan, Thomas D. Albright, and Gene R. Stoner, "Occlusion and the Interpretation of Visual Motion: Perceptual and Neuronal Effects of Context," *The Journal of Neuroscience* 20 (2000): 5885–5897.

binocular disparity cues[24]) such that they were perceived in different planes of depth. We used two complementary depth configurations: 1) foreground (*Near*) panels appear at the upper-right and lower-left corners of the stimulus and a background trough runs from the upper-left to lower-right (Figure 4B); and 2) foreground panels appear at the lower-right and upper-left and the background trough runs from lower-left to upper-right (Figure 4C). In both depth configurations, the diamond-shaped pattern of stripes appeared at an intermediate depth plane (above trough and below foreground panels). For each of these two configurations we moved the pattern of stripes either leftward or rightward within the stationary diamond-shaped window.

We predicted that the perceived direction of the stripes would vary depending on their depth relationships with the surround. To understand this prediction, note that the edges of the striped pattern that lie adjacent to a *Far* panel are readily (probabilistically) perceived as edges of an object in the visual scene. We call these "intrinsic" edges because they are intrinsic to the object that caused their presence in the retinal image. By contrast, edges of the striped pattern that lie adjacent to a *Near* panel are readily perceived as accidents of occlusion of one object by another. We call these "extrinsic" edges because they are extrinsic to all objects in the visual scene. Thus, the moving lines that rake (as retinal stimuli) along an intrinsic edge should be perceived as the consequence of a striped object moving through space in that direction. Conversely, the moving lines that rake along an extrinsic edge should be perceived as bearing no reliable relationship to the motion of the striped object, for it is impossible to verify, based upon these retinal motions, the extent to which the object moves under or parallel to the foreground surface. As a consequence of these depth-dependent perceptual interpretations of the cause of the retinal stimulus, the physically invariant motions should be perceived as arising from an object in the visual scene that slides along one or the other diagonal, depending on the spatial context. This is precisely what we found to be true; indeed, the effect is universal and striking.[25]

The configuration of these context-dependent motion stimuli enabled us to place them within the RFs of cortical neurons that respond selectively to the direction of stimulus motion. Such neurons are abundant in a small mid-level cortical region known as the *middle temporal visual area*, or *area MT*[26] (Figure 2). A typical MT neuron responds best to movement of a pattern in a specific direction – upward, for example – within the RF, and response wanes as direction deviates from this "preferred direction." Much evidence suggests that the activity of these neurons underlies the perception of motion,[27] so we reasoned that the response of an MT neuron to the barber-diamond stimulus should vary depending on the direction perceived (as dictated by spatial context), even though the motion in the retinal image never changes. Figure 5 illustrates data from an MT neuron that does this. Critically, only the moving stripes were placed within the RF of the neuron; the spatial context was visible to the RF surround. The response of this neuron, like many others in area MT (about 25 percent of the total population), is modulated by the spatial context such that it represents the perceptually inferred motion of an object in the visual scene, rather than the "sensory" motions in the retinal image. We naturally presume that the activity underlies the perceptual experience.

The barber-diamond provides a dramatic demonstration of how an invariant retinal stimulus can be perceived differently



Data were obtained by stimulating the RFs of neurons in area MT (see Figure 2) with unambiguously moving patterns of stripes (panel A) and with barber-diamond stimuli (panel B; see text and Figure 4 for stimulus description), for which the perceived direction is dictated by context.

(A) The striped pattern was moved in each of six different directions (indicated by small stimulus icons) within the RF of the recorded MT neuron. The motion percepts elicited by these six stimuli are determined solely by the physical motions of the stimulus. The mean neuronal responses are plotted in polar form, where the radius represents the frequency of action potentials (spikes per second) and the polar axis represents the direction that the stimulus moved. Like many MT neurons, this cell was highly selective, with a preference for motion down and to the right.

(B) The same MT neuron shown in (A) was in this case stimulated with four barber-diamond conditions. As detailed in text and in Figure 4, there are only two retinal image motion directions (left and right) present in these four stimuli, which are indicated by the black arrows on each stimulus icon. Because of the different depth configurations in the surround (see Figure 4), these two image motions are perceived to move in each of four unique directions (along the four diagonals), indicated by the gray arrows on each stimulus icon. We predicted that the neuronal response would be largest when the perceived motion matched the neuron's known direction preference (established from the test shown in panel A), such as down and to the right. This is precisely what was observed. Importantly, the retinal image motions for the two conditions on the right side of the panel are physically identical, but *the neuronal response reflects the direction perceived*, which is the visual scene motion inferred based upon spatial context.

Source: Robert O. Duncan, Thomas D. Albright, and Gene R. Stoner, "Occlusion and the Interpretation of Visual Motion: Perceptual and Neuronal Effects of Context," *The Journal of Neuroscience* 20 (2000): 5885–5897.

depending on context. Another property of perception is that it often generalizes across different *sensory attributes*[28]: widely varying retinal stimuli will be perceived as arising from a common source in the visual scene. These effects are termed *perceptual constancies*.[29] Size constancy, for example, refers to the fact that an object viewed from different distances will vary markedly in retinal image size but will be perceived as having the same size in the visual scene. This is a case in which the probable cause of the retinal stimulus is inferred using contextual cues for distance, such as linear perspective or binocular disparity.

In the laboratory we have studied a type of perceptual constancy known as "form-cue invariance," which occurs when essential attributes of the visual scene, such as

object form or motion, are extracted invariantly from the retinal image by generalizing across image attributes that distinguish object from background.[30] The profile of Abraham Lincoln's head, for example, is readily perceived regardless of whether it is defined relative to background by black-on-white, white-on-black, or red-on-green. This form of constancy reflects contextual information (evolutionary context) that is "built-in" the human visual system: to wit, we have evolved to operate in an environment in which the properties of light change over space and time, and objects are thus to varying degrees visible to us by different image cues, including brightness, color, and texture. Reliable perception depends on the ability to generalize across these state changes and extract the meaningful properties of the world.

We studied the neuronal basis of form-cue invariance using simple elongated rectangles that were defined by differences in brightness, texture, or flicker.[31] When these different stimuli are moved, the percept of motion is invariant, despite the fact that the sensory attributes of the image vary markedly. When we presented motion sensitive MT neurons with these form-cue invariant motion stimuli, we discovered that the neuronal responses also generalized across the different form cues. We concluded that these responses underlie the invariant perceptual experience by representing common motions of objects in the visual scene, to the exclusion of information about the sensory features that distinguish the objects from the background.

In the examples considered thus far, perception does not require reference to the observer's prior personal experience with the world. Much of the time, however, the way we perceive a sensory stimulus depends heavily on what we have seen before. This temporal context is manifested perceptually both as the ability to identify the cause of a sensory stimulus and by the ability to identify its meaning (as symbolic form). A simple but dramatic demonstration of such effects can be seen in Figure 6. Without prior experience with this particular pattern of retinal stimulation, it is difficult to identify a probable environmental cause (other than a surface with an apparently random pattern of light and dark regions) or to grasp what the sensory features symbolize. Figure 8 (page 35) provides information that will help disambiguate the image in Figure 6, leading to a coherent and meaningful percept. Similarly, the visual perceptual experience of *Devanagari* (Hindi script) is markedly different before and after learning the written language. These temporal context effects are, in fact, ubiquitous and fundamental to perceiving. They are rooted in the phenomena of *associative learning*[32] and *memory retrieval*[33]: as an observer learns the relationship between a sensory stimulus and the "farther facts associated with the object of sensation," the stimulus alone is capable of eliciting retrieval of those "farther facts," which become incorporated into the percept.

To understand the neuronal events that underlie the perceptual effects of temporal context, we have explored both the learning of sensory associations and the mechanisms by which associative recall influences perception. Sensory associative learning is the most common form of learning: it inevitably results when stimuli appear together in time, and particularly so in the presence of reinforcement; it can occur without awareness ("classical conditioning") or with ("instrumental conditioning"). The product of associative learning is that presentation of one stimulus elicits retrieval of its associate and all of the experience that that retrieval entails. For Pavlov's dog, for example, the learned association between the sound of

Figure 6
Demonstration of the Influence of Temporal Context on the Interpretation of a Sensory (Retinal) Stimulus

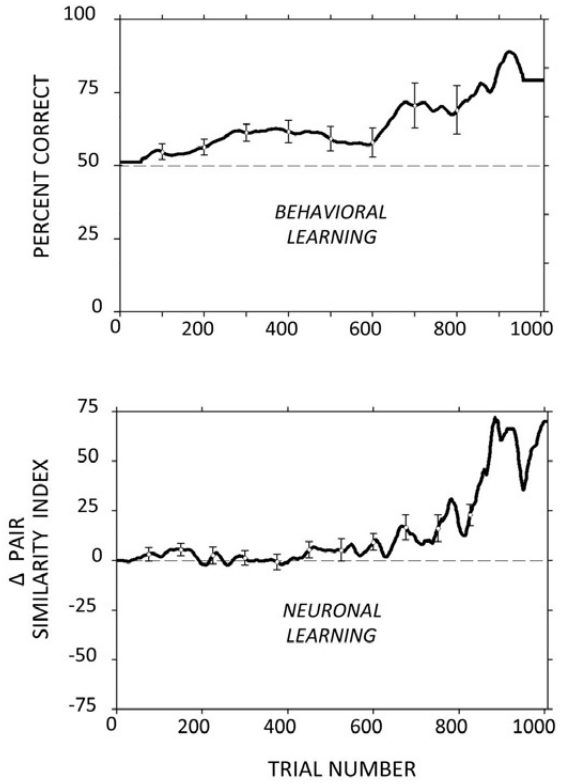
To most observers, this figure initially appears as a random pattern with no clear perceptual interpretation. The perceptual experience elicited by this stimulus is radically (and perhaps permanently) different after viewing the pattern shown in Figure 8. Source: Paul B. Porter, "Another Puzzle-Picture," *The American Journal of Psychology* 67 (3) (1954): 550 – 551.

the bell and the taste of food elicited in the dog the physiological manifestations of eating (such as salivating) when the bell alone was struck.

We began by examining how the learning of sensory associations is implemented by *neuronal plasticity*.[34] We hypothesized – in accordance with James's conjecture ("When two elementary brain-processes have been active together or in immediate succession, one of them, on reoccurring, tends to propagate its excitement into the other")[35] – that associations are established through the formation or strengthening of neuronal connections between the independent *neuronal representations*[36] of the paired stimuli. To test this hypothesis, we trained subjects to learn associations between pairs of visual stimuli that consisted of meaningless patterns ("clip-art" figures).[37] As the subjects acquired the associations, we monitored activity from neurons in the *inferior temporal (IT) cortex*,[38] a region that lies at the pinnacle of the *visual processing hierarchy*[39] (Figure 2) and is known to be critical for object recognition, which relies on memory of prior associations. Neurons in the IT cortex respond selectively to complex objects, such as the visual patterns used for this experiment.[40] We predicted that as subjects learned new associations between these stimuli, connectivity would increase between neurons that were initially selective for one or the other member of a pair. The result of this change in connectivity, then, should be expressed physiologically as a convergence of the magnitude of the neuronal responses to the paired stimuli. Figure 7 illustrates the average behavioral and neuronal changes that occurred in these experiments. As subjects learned novel visual stimulus associations (Figure 7, top panel), the index of neuronal response similarity (bottom panel) increased, reflecting the expected convergence of responses to paired stimuli. In other words, as the stimuli become symbols for one an-

Figure 7

Parallels between Behavioral Acquisition of Learned Associations between Pairs of Visual Stimuli (top panel) and Changes in the Selectivity of Cortical Neurons for the Paired Stimuli (bottom panel)

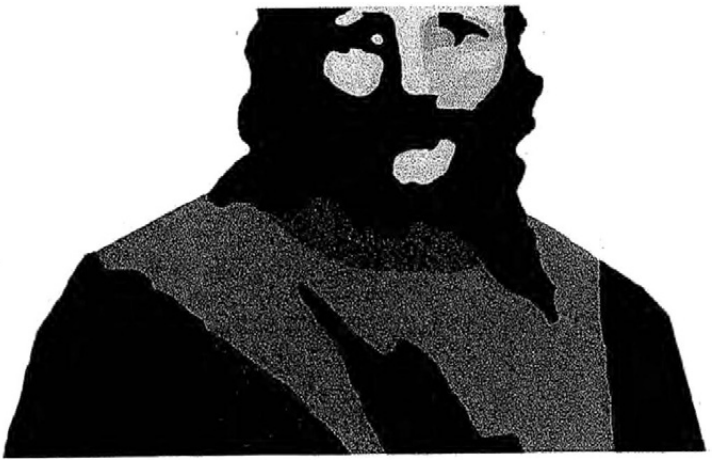


Top: Average performance as a function of trial number on a task in which subjects were instructed to choose a stimulus that we arbitrarily paired with a cue. For example, if subjects were shown stimulus A, they would be rewarded only for choosing stimulus B in the presence of distracters. Initial performance in this task is always at chance because the subject has no way to predict the pairings we impose. After a few dozen trials, performance begins to increase and ultimately reaches a high level.

Bottom: We predicted that concomitant with behavioral evidence of learning, we would see changes in the visual responses of neurons in the inferior temporal (IT) cortex (see Figure 2) to the paired stimuli. Plotted here is the trial-by-trial change in a measure of the similarity of response magnitudes to the paired stimuli (A and B). As learning proceeded, the responses to paired stimuli became more similar to one another and the responses to unpaired stimuli became less similar. We believe that these neuronal response changes reflect ongoing modifications to the local circuitry in the IT cortex, through which long-term memory for the learned pairings is achieved.

All data are averaged over many learning sessions, experimental subjects, and cortical neurons. Figure data: Adam Messinger, Larry R. Squire, Stuart M. Zola, and Thomas D. Albright, "Neuronal Representations of Stimulus Associations Develop in the Temporal Lobe during Learning," *Proceedings of the National Academy of Sciences* 98 (2001): 12239 – 12244.

Figure 8
Demonstration of the Influence of Temporal Context on the Interpretation of a Sensory (Retinal) Stimulus



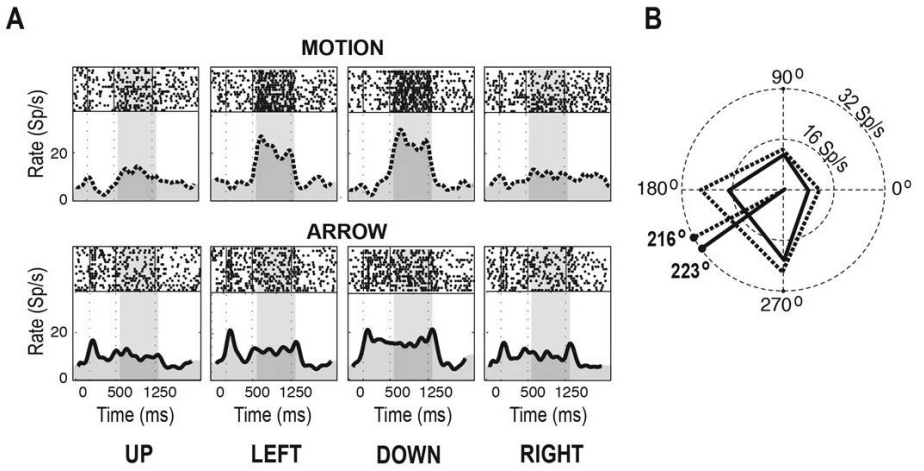
Most observers will experience a clear meaningful percept upon viewing this pattern. After achieving this percept, refer back to Figure 6. The experience of that image should now be markedly different, with a perceptual interpretation that is now driven largely by information drawn from memory. Figure previously published in Thomas D. Albright, "On the Perception of Probable Things: Neural Substrates of Associative Memory, Imagery and Perception," *Neuron* 74 (2) (2012): 227–245.

other by association, the neuronal representations of those stimuli become less distinguishable. We believe that these neuronal changes are the physical manifestations of the newly learned associative memories.

Memories consolidated[41] in this manner over the course of a lifetime provide the store of "farther facts associated with the object of sensation": the temporal context needed to interpret sensory events. We thus hypothesized that if the activity of a neuron in the visual cortex underlies perception (as opposed to sensation), that activity should be influenced by the retinal stimulus as well as by information retrieved from the *memory store* (the cellular locus of stored information in the brain). To test this hypothesis, we first trained subjects to associate pairs of stimuli and then evaluated the responses to individual members of each pair.[42] We predicted that the new "meaning" given to a visual stimulus by the learned association would be reflected in a new form of neuronal selectivity for the stimulus.

The stimuli used for this experiment were moving patterns and stationary arrows. Subjects learned that the direction of motion of each pattern was associated with the direction of the arrow. Upward motion, for example, was associated with an upwardly pointing arrow, and so on. We recorded activity from neurons in cortical visual area MT (Figure 2), which are highly selective for motion direction but are normally unresponsive to non-moving stimuli, such as the stationary arrows used in our experiment.[43] As expected, MT neurons responded selectively to the direction of pattern motion. After learning the motion/arrow associations, however, many MT neurons also responded selectively to the direction in which the stationary arrow pointed (Figure 9). More-

Perceiving *Figure 9*
Emergent Visual Stimulus Selectivity in Cortical Visual Area MT following Paired Association Learning



Subjects learned to associate up and down motions with up and down arrows. (A) Neurophysiological data are shown from a representative MT neuron. The top row illustrates responses to four motion directions. Each of the four plots in this row contains a spike raster display (in which each tic mark indicates the occurrence of an action potential) of the neuronal response obtained for each presentation of the stimulus. Each plot also contains a function that represents the average neuronal response rate as a function of time relative to stimulus onset (time = 0). For each plot, the vertical dashed lines correspond from left to right to stimulus onset, motion onset, and stimulus offset. The gray rectangle indicates the analysis window. The cell was highly directionally selective. The responses to downward and leftward motion were far greater than the responses to upward and rightward motion. The bottom row of plots illustrates responses to four static arrows. Plotting conventions are the same as in the upper row. As a consequence of learning, the cell became highly selective for arrow direction. Responses to the downward and leftward arrows are greatest.

(B) Mean responses of the neuron shown in panel A to motion directions (dashed curve) and corresponding static arrow directions (solid curve), indicated in polar format (polar angle corresponds to stimulus direction and radius corresponds to neuronal response rate). Preferred directions for the two stimulus types (dashed and solid vectors) are nearly identical.

Source: Anja Schlack and Thomas D. Albright, "Remembering Visual Motion: Neural Correlates of Associative Plasticity and Motion Recall in Cortical Area MT," *Neuron* 53 (2007): 881–890.

over, the selectivity seen was entirely predicted by the learned association. That is, if a cell responded best to upward motion, the arrow direction that elicited the largest response was also upward.

In these experiments, the associative training causes the arrow and motion to serve as symbols for one another in the same sense that the graphical pattern MOM serves as a symbol for a specific person. The perceptual experience elicited by sensing the arrow naturally includes the things that the arrow symbolizes, which in this case is the motion with which it has been paired. A corollary of James's axiom about "farther facts associated with the object of sensation" is that perceptual experience includes tangible "images" of the things recalled – images seen in the "mind's eye." Indeed, James defined his "general law of perception" as follows: "Whilst part of what we perceive comes through our senses from the object before us, another part (and it may be the larger

part) always comes out of our own head."[44] We thus infer that the perceptual experience elicited in our experiments by the arrow includes *visual imagery*[45] of the motion, which is retrieved from the memory store. We furthermore conclude that the selective pattern of activity exhibited by MT neurons upon viewing the stationary arrow underlies imagery of the motion recalled by association.

The conditions of the experiment described above elicit a form of visual imagery in which the thing imagined (motion) differs markedly from the sensory stimulus (arrow). This is the same form of imagery that occurs when we explicitly conjure the face of a friend, the characters and places in a novel, or imagine how the couch would look if we moved it to the other side of the room. But there is another form of imagery that is ubiquitous and largely nonvolitional, in which the thing (or things) recalled by association "matches" the sensory stimulus and is in fact a probabilistic inference about the cause of that stimulus. Harking back to our earlier discussion of the ambiguity of sensory events and the role of context in resolving that ambiguity, we can see that this latter form of imagery is a critical part of the process. This is particularly true under conditions in which the sensory stimulus is impoverished by noise or incompleteness, as is the case for the pattern in Figure 6. In fact, simple consideration of the objects that surround you – a chair that is partially obscured by a table, a glass blurred by glare, scratchy notes on a pad of paper – suggests that your clear and complete perceptual experience of these objects is the result of sensations that have been complemented – fleshed out, if you will – by information provided by memory. As our experiments have shown, that information comes in the form of selective feedback into the visual cortex, which unites retinal signals with imagery signals to yield perceptual experience of probable things.[46]

Given the ubiquitous, powerful, and implicit nature of this form of imagery, it is perhaps unsurprising that it can be readily manipulated to hijack perception. Indeed, a major category of performance magic relies on *priming*[47] as a form of temporal context. The result is an illusion in which the viewer's percept of an equivocal sensory event is not the "correct" solution. Similar unconscious priming effects occur under normal (non-magical) circumstances as well: perceptual biases in eyewitness reports or in the interpretation of wooly data (such as x-ray and forensic fingerprint examination and proofreading) are well documented. There are also priming effects on perception that happen with full awareness. *Pareidolia*[48] is the phenomenon in which we perceive coherent and meaningful patterns in response to (recognizably) random sensations, such as clouds that look like animals or foods that resemble Jesus. Finally, there are genres of art, such as impressionism, in which the image rendered – the sensory stimulus – is left intentionally vague in order to allow the perceptual experience to be completed by the viewer's prior experiences (which art historian E. H. Gombrich evocatively termed "the beholder's share").[49]

In this essay I have defined perceiving as a process by which the fundamental ambiguity of sensation is resolved through the use of contextual cues, which enable identification of the causes of sensation and attributions of meaning. Through examples drawn from our work on the visual cortex, I have illustrated how spatial context influences perception, how temporal context (memory) overcomes sensory noise and incompleteness, and how fundamental properties of the visual world (such as capricious illumination) are embedded in the machinery of our brains

(evolutionary context) and lead to perceptual constancies in the face of variable sensory conditions. From these discoveries, it is natural to conclude that a fundamental and generic computation in the cerebral cortex is the transformation from sensory attributes (the retinal image) to attributes of the external environment (the visual scene). While we know that context is used in that computation, and we can identify neuronal signals that reflect the outcome, we currently know very little about the neuronal mechanisms that give rise to these signals. How, for example, is information from memory selectively and dynamically routed back to the visual cortex in a context-dependent manner to complement information arising from the retina? How do neurons that represent visual motion incorporate information about the spatial layout of a visual scene? Identifying these processes presents formidable challenges, to say the least; but a variety of new experimental techniques – many from the fields of molecular biology and engineering – provide much promise for a future mechanistic understanding of perception.

ENDNOTES

1 Sensory transducers are specialized cellular mechanisms that enable various forms of environmental energy (mechanical, chemical, radiant) to be converted into energy that is communicated within and between neurons. Transduced energy leads to specific sensations. Photoreceptors, for example, transduce energy in the form of light into neuronal energy, which leads to visual sensation.

2 John Stuart Mill, *Examination of Sir William Hamilton's Philosophy and of the Principal Philosophical Questions Discussed in His Writings* (New York: Longmans, Green, Reader and Dyer, 1865).

3 William James, *The Principles of Psychology,* vol. 2 (New York: Henry Holt and Co., 1890).

4 James Sully, *Outlines of Psychology, with Special References to the Theory of Education* (New York: Appleton, 1888).

5 James, *The Principles of Psychology*.

6 Thomas Reid, *An Inquiry into the Human Mind, On the Principles of Common Sense*, ed. D. R. Brookes (University Park: Pennsylvania State University Press, 1764; 1997).

7 James, *The Principles of Psychology*.

8 The retina is a multilayered sheet of neuronal tissue that lines the back of the eye. Light refracted by the lens is projected onto the retina and is transduced by photoreceptors into neuronal signals. These neuronal signals are enhanced by additional retinal processing to detect contrast and are carried by the optic nerve/tract on to the rest of the brain.

9 James, *The Principles of Psychology*.

10 Ibid.

11 The cerebral cortex is a multilayered and highly convoluted sheet of neuronal tissue that forms the exterior of much of the mammalian brain. It mediates complex sensory, perceptual, and cognitive processes and coordinates voluntary goal-directed behaviors. The cerebral cortex makes up approximately three-quarters of the human brain's volume, and approximately one-third of the area of the human cerebral cortex is involved in the processing of visual signals.

12 Neurons are the major cell type in the brain involved in neuronal communication and computation. Signals are communicated electrically within a neuron via progressive changes in the voltage difference between the inside and outside of the cell. Signals are communicated chemically between neurons via the release of neurotransmitters into the synaptic cleft (the micro-

scopic space between neurons), which in turn activate receptors on the next neuron in the sequence. There are approximately one hundred billion neurons in the human brain.

13 The receptive field is the part of sensory space that, when stimulated, leads to a change in the activity of a sensory neuron (a neuronal response). In the case of vision, the receptive field of a neuron is defined by the region of visual space that, when stimulated, activates the neuron, and by the sensory attributes of the activating stimulus (such as its direction of motion).

14 An action potential is a brief stereotyped neuronal signal that typically sweeps from the cell body along the length of the output process (axon) of a neuron. This neuronal signal results from a rapid, active, and propagating exchange of ions across the cell membrane. An action potential entering the terminus of an axon leads to the release of the neurotransmitter.

15 The optic nerve/tract is a bundle of neuronal fibers composed of axons leaving the retina, which carry visual signals up to the rest of the brain. There are approximately one million retinal axons in each optic nerve of the human brain. The nerve becomes the optic tract when it enters the larger mass of the brain.

16 Synapse: 1) *Noun*: A collection of structures that underlies chemical transmission of signals between neurons. It includes specialized cell membranes on pre- and post-synaptic neurons, neurotransmitter substance, vesicles that package and release a neurotransmitter from a pre-synaptic membrane, receptors on a post-synaptic membrane that are activated by a neurotransmitter, and the synaptic cleft (space between pre- and post-synaptic membranes). 2) *Verb*: To form a synapse.

17 The forebrain is the most anterior portion of the vertebrate brain, which includes the cerebral cortex, thalamus, and hypothalamus and is responsible for sensory, perceptual, cognitive, and behavioral processes.

18 The lateral geniculate nucleus is a structure in the central portion of the mammalian brain that receives direct input from the retina and distributes visual signals to the cerebral cortex.

19 The primary visual cortex (area V1) is the portion of the cerebral cortex that serves as the entry point for visual signals ascending from the lateral geniculate nucleus of the thalamus. Area V1 lies on the posterior (occipital) pole of the human cerebral cortex (see Figure 2). Outputs from area V1 extend to a number of secondary and tertiary cortical visual areas in a hierarchical fashion.

20 Neurons represent and communicate information synthesized from their inputs in the form of a response, which is manifested and measureable as changes in the frequency (or pattern, in some cases) of action potentials.

21 For review, see Thomas D. Albright and Gene R. Stoner, "Contextual Influences on Visual Processing," *Annual Review of Neuroscience* 25 (2002): 339 – 379.

22 Gene R. Stoner, Thomas D. Albright, and Vilayanur S. Ramachandran, "Transparency and Coherence in Human Motion Perception," *Nature* 344 (1992): 153 – 155 ; Gene R. Stoner and Thomas D. Albright, "Neural Correlates of Perceptual Motion Coherence," *Nature* 358 (1992): 412 – 414 ; Gene R. Stoner and Thomas D. Albright, "The Interpretation of Visual Motion: Evidence for Surface Segmentation Mechanisms," *Vision Research* 36 (1996): 1291 – 1310; Robert O. Duncan, Thomas D. Albright, and Gene R. Stoner, "Occlusion and the Interpretation of Visual Motion: Perceptual and Neuronal Effects of Context," *The Journal of Neuroscience* 20 (2000): 5885 – 5897; Xin Huang, Thomas D. Albright, and Gene R. Stoner, "Adaptive Surround Modulation in Cortical Area MT," *Neuron* 53 (2007): 761 – 770; Xin Huang, Thomas D. Albright, and Gene R. Stoner, "Stimulus Dependency and Mechanisms of Surround Modulation in Cortical Area MT," *The Journal of Neuroscience* 28 (2008): 13889 – 13906; and Albright and Stoner, "Contextual Influences on Visual Processing."

23 Duncan, Albright, and Stoner, "Occlusion and the Interpretation of Visual Motion: Perceptual and Neuronal Effects of Context"; and Albright and Stoner, "Contextual Influences on Visual Processing."

24 Binocular disparity cues are subtle differences between the visual images cast in the two eyes that originate from the fact that right and left eyes have overlapping but slightly different

views of the world. These visual image differences ("cues") are physically related to the distance of a stimulus. The perception of distance in human vision is thus informed by these cues.

[25] Albright and Stoner, "Contextual Influences on Visual Processing"; see video demonstration of the effect at http://vcl-s.salk.edu/Research/Motion-Integration/.

[26] The middle temporal visual area (area MT) is the visual area of the cerebral cortex that receives direct input from area V1 and lies at an intermediate stage in the hierarchy of processing streams in the visual cortex (see Figure 2). Neurons in area MT are highly responsive to motion in the visual field and their activity underlies visual motion perception. Outputs of area MT control brain regions that mediate eye movements.

[27] For review, see Thomas D. Albright, "Cortical Processing of Visual Motion," in *Visual Motion and its Use in the Stabilization of Gaze*, ed. Joshua Wallman and Frederick A. Miles (Amsterdam: Elsevier, 1993), 177–201.

[28] Sensory attributes are basic elements of sensory experience given by properties of the sensory world. In the case of vision, the attributes of sensation include brightness, color, texture, distance, and motion, as well as size and shape. Combinations of sensory attributes elicit specific perceptual experiences.

[29] Perceptual constancy is the invariant perception of a visual object across changes in the sensory attributes that are manifested by the object. For example, perceived size is typically constant for a given object despite the fact that retinal size is inversely proportional to viewing distance. Viewing distance, in this case, is computed from visual depth cues, such as binocular disparity and linear perspective, and serves to normalize perceived size.

[30] Thomas D. Albright, "Form-Cue Invariant Motion Processing in Primate Visual Cortex," *Science* 255 (1992): 1141–1143; Gene R. Stoner and Thomas D. Albright, "Motion Coherency Rules are Form-Cue Invariant," *Vision Research* 32 (1992): 465–475; and Avi Chaudhuri and Thomas D. Albright, "Neuronal Responses to Edges Defined by Luminance vs. Temporal Texture in Macaque Area V1," *Visual Neuroscience* 14 (1997): 949–962.

[31] Albright, "Form-Cue Invariant Motion Processing in Primate Visual Cortex."

[32] Associative learning is the simplest and most common form of learning, in which different sensations, symbols, concepts, or events are linked together in memory by virtue of their proximity in time and/or reinforcement.

[33] Memory retrieval is the process of accessing information stored in memory and making it available for use in perceptual, cognitive, and behavioral processes.

[34] In this context, plasticity refers to experience-dependent changes in the stimulus attributes that are encoded by a neuron.

[35] Albright, "Form-Cue Invariant Motion Processing in Primate Visual Cortex."

[36] The unique stimulus attributes that lead to a response from a neuron (a change in neuronal activity) define the information that is symbolically represented by the neuron. For example, a visual neuron that responds selectively to a particular direction of motion in its receptive field is said to "represent" that direction of motion.

[37] Adam Messinger, Larry R. Squire, Stuart M. Zola, and Thomas D. Albright, "Neuronal Representations of Stimulus Associations Develop in the Temporal Lobe during Learning," *Proceedings of the National Academy of Sciences* 98 (2001): 12239–12244.

[38] The inferior temporal (IT) cortex is the area of the cerebral cortex that corresponds to the highest stage of purely visual processing. The IT cortex is located on the inferior convexity of the temporal lobe in primates (see Figure 2) and plays a central role in visual object recognition. Neurons in the IT cortex represent complex collections of stimulus attributes that correspond to behaviorally meaningful objects, such as faces. IT neurons exhibit experience-dependent changes in their receptive field properties during visual associative learning, and those changes constitute the cellular trace of the associative memory.

39 The visual processing hierarchy is a succession of visual processing stages leading, for example, from the retina, through the thalamus, to the primary visual cortex, and on to secondary and tertiary cortical visual areas. Each stage in the hierarchy selectively integrates information from the preceding stage, yielding increasingly complex neuronal representations as signals move up through the hierarchy.

40 Robert Desimone, Thomas D. Albright, Charles G. Gross, and Charles J. Bruce, "Stimulus Selective Properties of Inferior Temporal Neurons in the Macaque," *The Journal of Neuroscience* 8 (1984): 2051 – 2062.

41 Memories are initially encoded in a form that has limited duration and capacity and is labile, decaying quickly with time and easily disrupted by other perceptual or cognitive processes. Through cellular and molecular events that play out over time, the contents of short-term memories may be encoded and consolidated into long-term memory, which is more enduring and of greater capacity.

42 Anja Schlack and Thomas D. Albright, "Remembering Visual Motion: Neural Correlates of Associative Plasticity and Motion Recall in Cortical Area MT," *Neuron* 53 (2007): 881 – 890.

43 Albright, "Cortical Processing of Visual Motion."

44 James, *The Principles of Psychology*.

45 Visual imagery is the subjective experience that occurs when memories of perceptual experiences are retrieved. Imagery of an object is similar to a percept of the object resulting from retinal stimulation; both are mediated by the same neuronal structures and events in the visual cortex. Explicit imagery is volitional (for example, picture your car in your mind's eye). Implicit imagery is cued by association with a sensory stimulus and serves to augment perceptual experience based on expectations, in the face of noisy or incomplete sensory signals.

46 For review, see Thomas D. Albright, "On the Perception of Probable Things: Neural Substrates of Associative Memory, Imagery, and Perception," *Neuron* 74 (2012): 227 – 245.

47 After two sensory stimuli have become associated with one another, presentation of one stimulus leads to enhanced (faster, more accurate) processing of the other. This phenomenon is known as priming.

48 Pareidolia is a phenomenon of visual imagery in which expectations derived from prior experience cause random patterns to be perceived as meaningful objects.

49 Ernst H. Gombrich, *Art and Illusion: A Study in the Psychology of Pictorial Representation* (Princeton, N.J.: Princeton University Press, 1961).

The Energetic Ear

A. J. Hudspeth

Abstract: As the gateway to human communication, the sense of hearing is of enormous importance in our lives. Research on hearing has recently been revolutionized by the demonstration that the ear is not simply a passive receiver for sound, but also an amplifier that augments, filters, and compresses its inputs. Hair cells, the ear's sensory receptors, use two distinct methods to implement an active process that endows our hearing with these remarkable properties. First, the vibration-sensitive structures of the ear, called hair bundles, display a mechanical instability that allows them to oscillate in response to stimulation. And second, the membranes of hair cells are replete with proteins that contract in response to electrical stimuli, thus enabling the cells to act like tiny muscles. The activity of these two motile processes can be so exuberant as to cause normal ears to emit sounds.

A. J. HUDSPETH, a Fellow of the American Academy since 2002, is the F. M. Kirby Professor and Head of the Laboratory of Sensory Neuroscience at the Rockefeller University. He is also an Investigator at the Howard Hughes Medical Institute and a member of the National Academy of Sciences. His research has been published in journals such as *The Journal of Neuroscience*, *Nature Reviews Neuroscience*, *Nature*, and *Neuron*; he recently coedited the fifth edition of the textbook *Principles of Neural Science* (2013).

Most of us use hearing aids. Electronic devices amplify sound for the benefit of those with compromised hearing, whereas the ears of people with normal hearing contain biological structures that serve an identical function. This so-called active process can amplify sounds by more than a hundredfold. The ear's intrinsic amplifier additionally tunes our responsiveness to specific frequencies of sound, thus facilitating the recognition of sound sources and the discrimination of speech. The active process also allows us to analyze acoustic signals over a million-fold range of magnitudes, compressing responses so that we can appreciate both soloist and orchestra. Most remarkably, the ear's native hearing aid can, like an electronic device, become unstable, leading to the emission of sounds from the ear. Hearing is a highly adaptive sensory modality, for it provides an early warning of potential adversaries or predators and a foretaste of possible prey while they are still distant. In these endeavors, there is a clear advantage in having the most sensitive and discriminating auditory apparatus. Evolution has accordingly fostered an active process whose performance approaches the limits set by the physics of sound.

doi:10.1162/DAED_a_00316

Because hearing is the key sense in human communication, its importance is most apparent in persons whose hearing is deficient. Hearing is ordinarily the means by which children acquire language and thus their avenue to other forms of symbolic communication. One child in a thousand is born deaf, however, and a comparable percentage will become deaf before maturity, largely a result of the several hundred forms of genetic hearing loss.[1] Before the advent of a simple and effective test to identify these children within days of birth, many suffered years of retarded development owing to their unrecognized condition.

Verbal exchanges not only facilitate the transmission of information but also situate us in our social milieux, yet forty million Americans – an eighth of the populace – have hearing loss severe enough to mar their daily lives, for example by impeding conversation on the telephone or in a noisy environment. Age-related hearing loss, or presbycusis, affects 25 percent of our population at age sixty-five and 50 percent by age eighty, distancing many individuals from friends and family and greatly diminishing their quality of life. Finally, abrupt hearing loss – whether from overloud sounds, infections, or certain medications – may afflict individuals of any age. Deafness can be psychologically devastating, leading to depression and even suicide as a result of the isolation that it imposes. As Helen Keller remarked in a letter:

> I am just as deaf as I am blind. The problems of deafness are deeper and more complex, if not more important, than those of blindness. Deafness is a much worse misfortune. For it means the loss of the most vital stimulus – the sound of the voice that brings language, sets thoughts astir and keeps us in the intellectual company of man.[2]

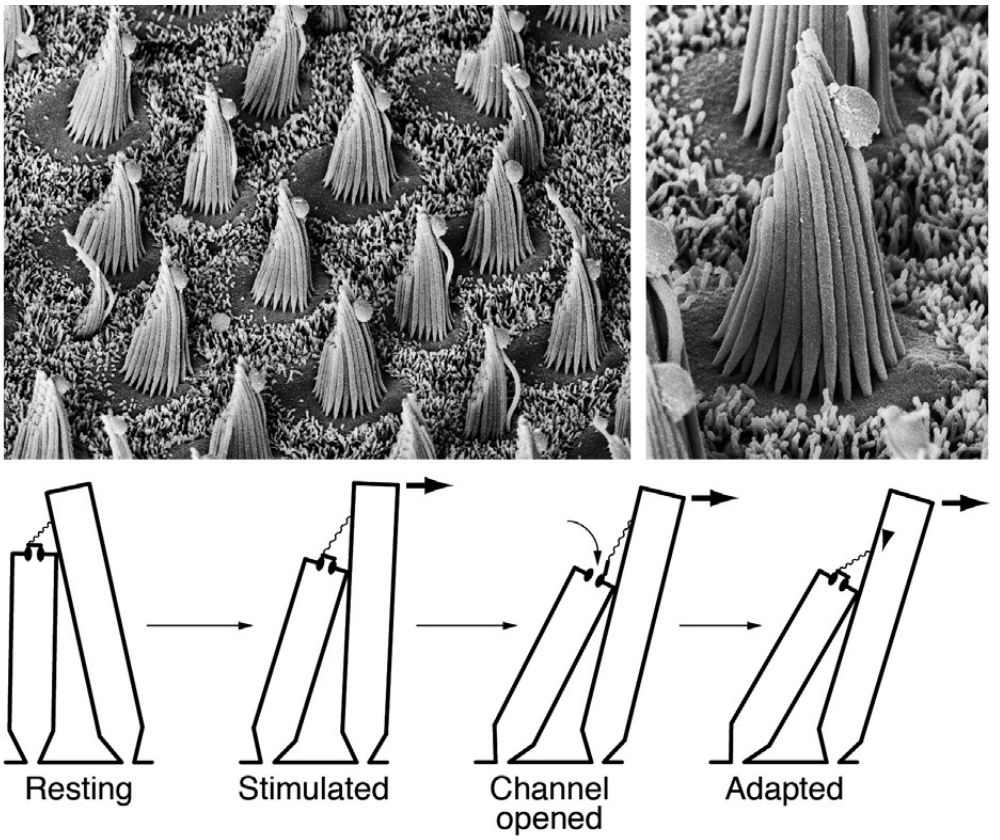
A. J. Hudspeth

The basic outline of the hearing process is familiar to us from high school biology.[3] Sound waves – the alternate compressions and rarefactions of the air associated with a sound – are captured by the external ear, traverse the ear canal, and strike the thin, elastic eardrum. The ensuing vibrations propagate through the three miniscule bones of the middle ear – the hammer, anvil, and stirrup – to the cochlea. Named for its resemblance to a spiraled snail shell (*κόχλος* in Greek), the cochlea encompasses three helical turns in an organ the size of a chickpea. Abutting the cochlea within the temporal bone of the skull are the five receptor organs of the vestibular labyrinth, our source of information about rotatory movements and linear accelerations including gravity.

The cochlea comprises three tiny, liquid-filled tubes that spiral in parallel from the organ's base to its apex. Vibration of the stirrup bone applies an alternating pressure to the contents of one chamber, setting in motion the elastic boundaries between the tubes. One of these boundaries, the basilar membrane, supports the mechanically sensitive structure of the ear, the organ of Corti. Here the vibration occasioned by the incoming sound is reinterpreted as an electrical signal, the common currency of signaling throughout the nervous system. This process, the analog of detecting light in the eye or an odorant in the nose, constitutes auditory transduction.

The receptors responsible for transduction are termed hair cells, for each bears on its top surface a mechanically sensitive organelle called the hair bundle, which comprises between ten and three hundred regularly spaced, erect, cylindrical protrusions called stereocilia (Figure 1). This name, which means "stiff hairs," characterizes an important feature: each stereocilium contains a rigid fascicle of cross-linked filaments of the protein actin. When a mechanical force is applied at the top of the

Figure 1
The Mechanotransduction Process of Hair Cells



The scanning electron micrograph at the upper left shows the surface of the sacculus, a sensory organ in the inner ear of a frog. Conical hair bundles extend about three ten-thousandths of an inch from the smooth tops of the mechanically sensitive hair cells. The hair cells are separated by supporting cells marked by a stubble of fine protrusions called microvilli. The enlargement at the upper right portrays a single hair bundle comprising about sixty stereocilia and a lone kinocilium with a bulbous tip at the tall edge of the bundle. Note the progressive increase in stereociliary length from left to right; deflecting the bundle in the same direction excites the cell.

The diagram of two adjacent stereocilia schematizes the mechanism of transduction. When the hair bundle stands at rest, the filamentous tip link interconnecting the stereocilia bears little tension and the channel at its lower end is usually closed. An excitatory stimulus (thick arrow) deflects the bundle toward its tall edge, causing a sliding motion between the stereocilia and consequently increasing the tension in the link. This tension opens the channel, allowing positively charged ions to carry electrical current into the cell (curved arrow). If the stimulus persists for more than a few hundredths of a second, the hair cell adapts: the upper insertion of the tip link slides down the longer stereocilium (arrowhead), relaxing the tip link and allowing the channel to reclose. Note that the relative proportions in the diagram have been exaggerated in the interest of clarity. In reality, the loudest tolerable stimulus would deflect the hair bundle by only one-quarter of a stereociliary diameter and the channel would be smaller than the line thickness. Source: Figure prepared by the author.

hair bundle, each constituent stereocilium pivots at its tapered base with minimal flexing along its shaft. This movement entails a sliding motion between adjacent stereocilia that is central to the transduction process.

The stereocilia in a hair bundle are not of uniform length. A poorly understood developmental process causes one row of these stereocilia to grow longest, while also rendering each successive row progressively shorter. Every hair bundle is accordingly beveled like the tip of a hypodermic needle. This characteristic is of importance in transduction, for the bundle is most sensitive to deflection along the direction of its bevel. Displacement of the hair bundle's top toward its tall edge excites the hair cell, whereas motion in the opposite direction has an inhibitory effect. Movement at a right angle, perpendicular to the bundle's plane of mirror symmetry, elicits no response. The hair bundle can thus be considered a biological strain gauge that is excited or inhibited by appropriately oriented mechanical stimuli.

Like other excitable cells, the hair cell produces electrical signals across its surface membrane through the action of ion channels, which are proteins that traverse the membrane and offer tiny pores through which electrically charged ions can flow. Most channels are equipped with some form of molecular gate that can open or close to regulate the flux of ions. The channels responsible for signaling in the nervous system, for example, include those responsive to membrane voltage, which underlie the propagating signals called action potentials, and those sensitive to neurotransmitter chemicals, which mediate the synaptic interactions between neurons. The ion channels of the hair bundle instead have mechanically sensitive gates that open and close in response to bundle displacements.

Stimuli are conveyed to the mechano-electrical-transduction channels by tip links, which are molecular threads consisting of four molecules of extracellular proteins called cadherins that run obliquely from the tip of each stereocilium to the flank of the longest adjacent stereocilium (Figure 1). Each of these tip links is likely connected to two ion channels at its lower end. When the top of a hair bundle is pushed toward its tall edge, the sliding between adjacent stereocilia increases the tension in the obliquely mounted link, which pulls the channels open. Movement in the opposite direction allows a small fraction of the channels that are open at rest to close. Because the tip links are oriented parallel to a bundle's bevel, they do not sense perpendicular stimuli.

The direct linkage of channel opening to hair-bundle movement has three important consequences for our hearing. First, because there are no chemical intermediates in the transduction process, hearing is rapid. We are able to detect sounds at frequencies as great as twenty thousand oscillations per second, and bats and whales can respond to stimulation at frequencies five times that great. By comparison, we process visual stimuli at only one-thousandth of that speed: in motion pictures and television, images presented twenty or thirty times per second are construed as continuous because they exceed the eye's rate of transduction.

The direct opening of channels next explains why hearing is so sensitive. The faintest sounds that we can perceive vibrate hair bundles by ten billionths of an inch, which is an atomic dimension. This movement is analogous to the apex of the Washington Monument budging by less than an inch. The sensitivity of hearing is limited by noise produced by the water molecules of the inner ear's fluids that, excited by heat, continually buffet hair bundles.

The Energetic Ear

The third and most important consequence of the direct mechanical gating of transduction channels is the mechanical instability of the hair bundle. Applying a small force to a hair bundle tenses the tip links, which may in turn activate several transduction channels. Because opening the channels' molecular gates relaxes the attached tip links, the links associated with the remaining channels must bear more of the stimulus force. If these channels then open as a consequence, an avalanche ensues: the activation of a few channels triggers the opening of the rest. Similar behavior occurs when more than a certain number of channels have been shut by stimulation in the opposite direction; now the remaining channels close in concert. The consequence is that a hair bundle becomes bistable, adopting either a configuration with open channels upon stimulation toward its tall edge or a state with closed channels in response to force in the opposite direction. The bundle cannot, however, remain stably in a position between the two extremes. This behavior resembles that of a child's click toy: a metal strip that snaps between two shapes with an audible pop. As we shall see, the instability of the hair bundle has been harnessed by evolution to implement the active process.

Sounds can be captured effectively through the phenomenon of resonance, in which each cycle of a stimulus tone adds a tiny increment of energy to a vibrating structure. This is how an opera singer's voice can shatter a champagne glass: prolonged vocalization of the note to which the glass is tuned causes an oscillation that grows in amplitude until the material fails. Hearing would be most sensitive if the ear's structures were free to accumulate sound energy through resonance; each individual hair bundle would vibrate at a specific frequency determined by its dimensions and material properties. Doing this would present a challenge, however, because a hair bundle, like other cellular components, requires a liquid environment. Movement through the extracellular fluid causes a sound wave to dissipate, losing power due to viscosity, which reflects the friction between a hair bundle and the surrounding water molecules. The active process represents evolution's reconciliation of these issues. By continually supplying energy to the vibrating hair bundles, the active process counters the energy-dissipating effects of hydrodynamic friction and allows our hearing to exploit resonance.

As its name suggests, the *active* process must work to overcome viscosity and amplify a hair bundle's mechanical inputs.[4] The law of conservation of energy implies that a hair cell must draw on some source of biochemical energy to power its active process. At least in the ears of land-dwelling, non-mammalian vertebrates, a form of myosin – the type of protein that animates our muscles – does the work that underlies the active process. Rather like the participants in a tug o' war, myosin molecules consume cellular energy to pull against intracellular strands of the protein actin. Through such exertions, a cluster of myosin molecules at the upper end of each tip link continually tightens the link and thus ensures that some transduction channels remain open. If movement of the hair bundle toward its tall edge further tenses the link and activates more channels, then the myosin molecules work less and allow the link to relax and some channels to reclose. Conversely, when a negative hair-bundle movement slackens the link and closes the channels, the myosin molecules ascend, restoring tension and thereby reopening the channels. This process is called adaptation: whenever a protracted displacement is applied to the bundle, transduction channels transiently

open or close, after which the myosin movements restore the *status quo ante* within a few hundredths of a second. The adaptation process ensures that the machinery of transduction is always poised where it is most effective: on the brink of channel opening.

Imagine that a hair bundle is somehow deflected toward its tall edge, perhaps by collision with an energetic water molecule, such that most of the channels snap open. The myosin molecules react by lowering the tip-link tension in an effort to restore the probability of channel opening to an equilibrium value of one-half. Before that point is reached, however, the closure of some channels triggers an avalanche in the opposite direction and most of the remaining channels snap shut as well. In response, the myosin molecules begin to ascend in a renewed effort to achieve equilibrium. But again they are thwarted, for as the rising tip-link tension opens some channels, the remainders stampede in the same direction. As the system jumps back and forth in a futile attempt to achieve equilibrium, the hair bundle oscillates from side to side. Mechanical recordings from hair bundles have demonstrated these spontaneous movements *in vitro*.[5]

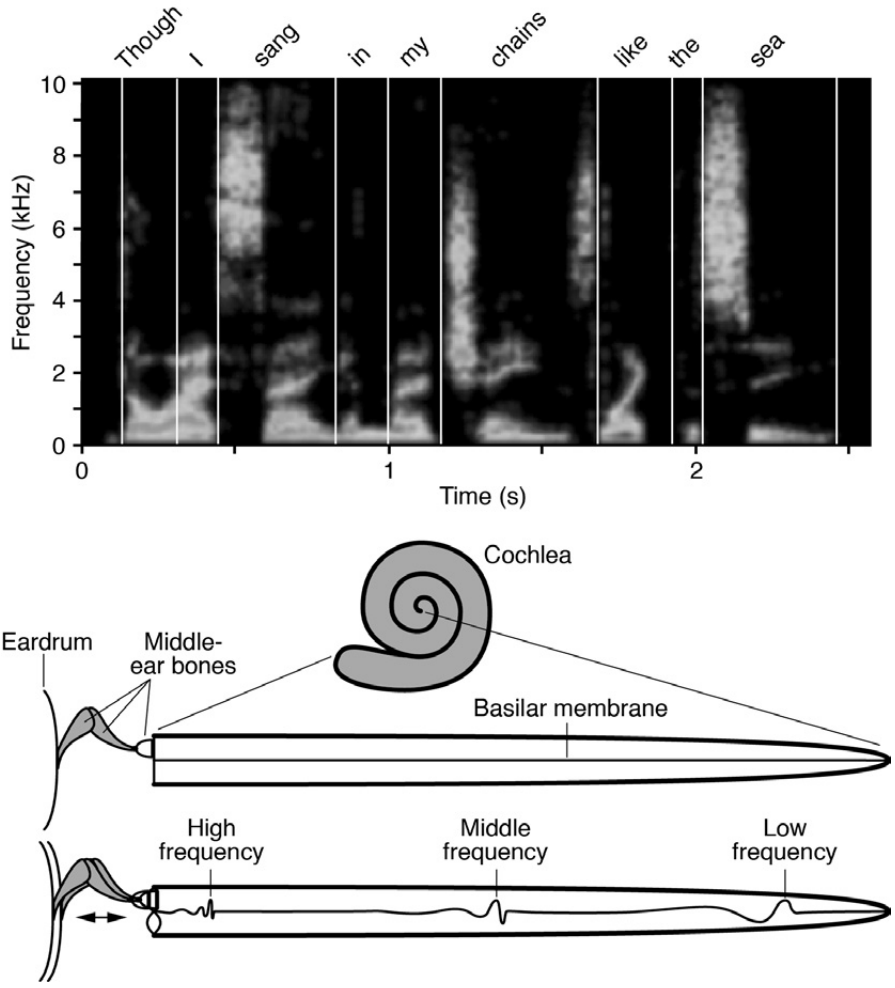
A mechanical stimulus applied to an oscillatory hair bundle harnesses this activity. If the frequency of stimulation is significantly greater or less than that at which the hair bundle oscillates spontaneously, then the ensuing response is small. If the stimulus accords with the bundle's natural frequency of oscillation, however, the adaptation process can pump an increment of energy into each cycle of movement. Just as the motion of a child's swing gradually increases with each parental push, the response grows over a few cycles to a peak amplitude at least a hundred times larger than that of a passive system. In the extensively studied ear of the frog, active hair-bundle motility accounts quantitatively for the active process. Bundle motility contributes to the active process in mammals as well, but, as discussed below, in these animals it is supplemented by an additional mechanism.

Among the most striking and useful features of hearing is our ability to distinguish tones of different frequencies. Although a semitone, the narrowest interval on a piano, represents a frequency difference of about 6 percent, a trained listener can readily detect an interval of only 0.2 percent – in musical parlance, three cents, or hundredths of a semitone. Frequency discrimination is of obvious importance in the performance and appreciation of music. However, this faculty is also used continually in daily life in our recognition of sound sources and our interpretation of speech. The distinctions between various phonemes, or speech sounds, rest upon the frequencies present in each (Figure 2). To identify a speaker and, more particularly, to determine what she has said, the cochlea must somehow parse complex sounds into their constituent frequencies.

In 1863, German physician and physicist Hermann von Helmholtz first appreciated that the cochlea works like an inverse piano. A piano blends the sounds from several independent resonators – the oscillating strings – into a harmonious whole. The cochlea undoes this effort, separating from a complex sound the various pure tonal constituents. Every component is then analyzed independently so that the brain receives a description of each successive phoneme in terms of its constituent frequencies. From this information, the central nervous system can calculate what word was spoken while also abstracting such nuances as accent and emotional inflection.

How are the different components of a complex sound separated? This operation

Figure 2
The Operation of the Mammalian Cochlea



The sonogram in the upper panel analyzes my voice declaiming the final line of Dylan Thomas's poem "Fern Hill."[6] Time (in seconds) is represented on the horizontal axis and frequency (in thousands of cycles per second) on the vertical; loudness is signified by brightness. The prominent vowels in "though," "sang," "my," "chains," and "sea" involve multiple low frequencies – the formants – whereas the consonants are represented by high frequencies. When responding to speech such as this, the cochlea rapidly and continuously deconstructs phonemes into their constituent frequencies; the brain then uses the resultant information to infer what has been heard.

The schematic drawings in the lower panel show how the snail-shaped cochlea would look if unrolled: a long, bone-enclosed tube bisected by the elastic basilar membrane. Sound striking the eardrum sets it and the three tiny bones of the middle ear into oscillation, which is communicated to the liquid contents of the cochlea. Each frequency component of the sound elicits a traveling wave that propagates along the basilar membrane and peaks at a specific position corresponding to that frequency. By this means, the cochlea separates complex sounds into vibrations that stimulate certain of the sixteen thousand hair cells arrayed along the basilar membrane. Note that the vertical motion of the membrane has been exaggerated three hundred thousand–fold with respect to the cochlea's length. Source: Figure prepared by the author.

is performed by the basilar membrane, one of the elastic boundaries of the cochlear chambers that are deflected by sound pressure. The basilar membrane extends about one and one-third inches along the cochlear spiral and varies continuously in its physical attributes. At the base of the cochlea the membrane is narrow, light, and taut; like the thinnest string on a violin, it oscillates at a high frequency. The basilar membrane at the cochlear apex, which like the coarsest string on a contrabass is broad, massive, and floppy, instead resonates upon stimulation at a low frequency. Each intermediate position is most responsive to stimulation at a specific frequency that grows incrementally from the apex, which responds to twenty vibrations per second, to the base, which is sensitive to twenty *thousand* cycles a second.

When a particular tone is sounded, the basilar membrane begins to vibrate at the same frequency. The membrane does not move as a unit, though; instead, the oscillation propagates from the base toward the apex as a traveling wave (Figure 2). The motion is small at the membrane's base but grows progressively larger as the wave approaches the place that responds to that specific frequency. There the wave achieves its peak amplitude before rapidly dissipating like a comber breaking upon a beach.

A complex sound engenders overlapping but largely independent traveling waves for each of its components. The position at which each wave peaks specifies its frequency; the magnitude of each peak indicates the wave's loudness. Arrayed in single file along the basilar membrane, four thousand receptors called inner hair cells detect these vibrations and produce electrical responses that are then conveyed to the brain by thirty thousand nerve fibers. The cochlea thus acts as a frequency analyzer, providing the central nervous system with a nearly instantaneous report of the tones present in any acoustic stimulus.

Because the basilar membrane oscillates within the liquids that fill the cochlear chambers, it too loses energy to viscosity. Research conducted during the past two decades has demonstrated how the active process counters this problem. In the mammalian cochlea, the active process is facilitated by twelve thousand outer hair cells that spiral along the basilar membrane in three parallel rows adjacent to the single row of inner hair cells. The outer hair cells transmit a negligible amount of information to the brain; instead, they serve as dedicated amplifiers that enhance the mechanical stimuli delivered to the inner hair cells. When sound displaces the basilar membrane, side-to-side movements of the tectorial membrane tweak the underlying hair bundles, exciting active hair-bundle motility like that of other hair cells. In addition, transduction of the mechanical input elicits in outer hair cells an electrical response that drives a unique phenomenon called *somatic motility*. When the cell's voltage becomes more negative, the millions of prestin molecules studding the membrane of each outer hair cell expand; a more positive voltage causes contraction. The entire cell consequently changes in length, respectively elongating or shortening like a tiny muscle. And like a muscular contraction, somatic motility delivers energy, accentuating the basilar membrane's vibrations. This activity constitutes an example of positive feedback in which sound-induced vibrations beget still larger oscillations. And in the active process, as in a public-address system, positive feedback can lead to instability.

For decades after the first proposal that the ear employs an active process, there was little evidence to support the hypothesis. This changed when careful measurements showed that, in a very quiet sound chamber, the ears of at least 70 percent of normal humans emit one or more tones. In other

words, the ear not only takes sound in, but also puts sound out! These spontaneous otoacoustic emissions are not pathological, but on the contrary are a hallmark of healthy ears whose hair cells are capable of exuberant activity.

As discussed above, the active process is an example of positive feedback. Moreover, like many man-made feedback systems, the active process exhibits gain control: it can be turned up or down as circumstances dictate. When amplification is unnecessary in a loud environment, the feedback largely vanishes and the ear is essentially passive. Under the conditions that prevail through most of daily life, corresponding to sound-pressure levels of approximately sixty decibels, amplification makes only a modest contribution to the ear's responsiveness. Near the threshold of hearing around zero decibels, however, we feel that we can hear a pin drop: indeed, our acoustic sensitivity is enhanced more than a hundredfold by the active process operating at its highest gain. An individual who lacks the active process as a consequence of hair-cell damage loses this amplification and therefore becomes hard of hearing.

What happens if the strength of feedback increases still further? The result is called a bifurcation: an abrupt, qualitative change in the behavior of the auditory apparatus. Just as slowly turning up a public-address system suddenly elicits a howling noise, the ear can reach the point at which some hair cells begin to oscillate spontaneously. The vibrations are transmitted back out of the cochlea, resulting in spontaneous otoacoustic emissions that can be measured in the ear canal. In rare cases, these sounds can be heard by nearby people, even at a distance of several inches. There can be no doubt in such instances that the cochlear amplifier is active, for the ear radiates energy into the surrounding air.

Spontaneous otoacoustic emissions are not the only unexpected emanations of a normal human ear. The instability in hair-cell transduction also generates combination tones, sounds that are heard and even emitted from ears though they are absent from acoustic stimuli. Suppose one listens to simultaneous, moderately loud sounds of two distinct but nearby pitches: a higher frequency f_2 and a lower one f_1. In addition to these two tones, a normal listener then hears the prominent combination tone $2 \cdot f_1 - f_2$, twice the lower frequency minus the higher. This tone is somewhat fainter than the two that are actually sounded, but is nevertheless loud enough to be perceived clearly. In fact, composers have created music in which no instrument actually plays the audible melody. Two streams of tones are played instead, and the listener's ear synthesizes the melody from the successive tone pairs. Composers György Ligeti and Karlheinz Stockhausen are among those who have experimented with this effect, which was first discovered in the eighteenth century by violinist and composer Giuseppe Tartini.[7]

Because combination tones originate from the instability of normally functioning hair bundles, they provide a useful assay for normal hearing. Most newborn children in the United States are now subjected to a simple test in which two tones are played into each ear while a sensitive microphone records the sounds in the ear canal. If the measured intensity of the combination tone exceeds a certain threshold, hearing in that ear is almost certainly normal. If the response is diminished, however, other tests can confirm whether hearing is impaired and how it might be remedied.

Hearing loss is a growing problem in a society characterized by increases in both lifespan and noise pollution. Although cell

division continually replenishes most cells in the human body, a few critical types of cell – including hair cells in the ear as well as nerve cells in the brain and muscle cells in the heart – are unfortunately not replaced by this means. As we lose hair cells, we also forfeit the advantages afforded by the cochlea's active process. The ear's sensitivity thereby declines, rendering us hard of hearing. A diminished capacity to distinguish frequencies impairs our ability to recognize the subtle nuances of speech. And loss of the cochlea's compressive quality means that weak sounds become inaudibly soft and strong sounds offensively loud. It is for this reason that a hearing aid is so often unsatisfactory: the device intensifies the sounds reaching the ear but cannot restore the normal sharpness and dynamic range of hearing.

American Sign Language (ASL) provides one successful means of communication for the hearing-impaired. Used by half a million people in the United States and Canada, this form of signing represents a highly evolved language quite distinct from spoken English. Regional derivatives of ASL and other distinctive sign languages are used throughout the world. During the last few decades, the deaf have also benefitted from the introduction and evolution of the cochlear prosthesis. Surgically implanted into a damaged ear, this array of electrodes restores a degree of hearing by directly stimulating the nerve fibers that run from the cochlea into the brain. A receiver worn outside the head replaces the lost hair cells by deconstructing sounds into their component frequencies and sending electrical signals to the corresponding electrodes. As the most successful neural prosthesis to date, with nearly three hundred thousand users worldwide, the cochlear implant has raised hopes for progress with future electrode systems that might be used to restore vision or overcome spinal injuries.

The most enticing avenue for a long-term solution to deafness is through the regeneration of hair cells. Although fishes, amphibians, and reptiles (including birds) can readily replace these receptors, mammals for unknown reasons cannot. Numerous researchers are now attacking this problem, trying to understand how regeneration occurs naturally and why it is deficient in mammals. Although the pace of research is painfully slow for those with impaired hearing, recent results are encouraging. Various treatments have created mammalian hair cells *in vitro* and even in the damaged ears of animals, and investigators have begun to identify the molecular signals that underlie the decision of precursor cells in the ear to multiply and assume the role of hair cells.

The active process provides a striking example of the opportunistic nature of evolution. The direct mechanical gating of transduction channels – the simplest mechanism that might be envisioned – inevitably inflicts the distortion responsible for combination tones. This mode of action additionally imposes mechanical instability on the hair bundle. Despite these flaws, direct channel gating has apparently persisted by virtue of its great speed. The necessity of maintaining the transduction machinery within its narrow working range likely supported the evolution of adaptation. This process too has been implemented in a simple manner, through the activity of a common form of myosin, the workhorse of force production in cells. Yet most remarkably, combining the phenomena of direct transduction and adaptation has yielded the fundamental features of the active process: a tuned amplifier with a broad dynamic range, the foundation of our extraordinary hearing.

ENDNOTES

Author's Note: The author thanks Ms. Christina Black, Mr. Jeff Robinson, and Dr. Maurine Packard for comments on the manuscript.

1 Guy P. Richardson, Jacques Boutet de Monvel, and Christine Petit, "How the Genetics of Deafness Illuminates Auditory Physiology," *Annual Review of Physiology* 73 (2011): 311 – 334.

2 James Kerr Love, ed., *Helen Keller in Scotland: A Personal Record Written by Herself* (London: Methuen & Company Ltd., 1933), 68.

3 A. J. Hudspeth, "The Inner Ear," in *Principles of Neural Science*, 5th ed., ed. Eric R. Kandel, James H. Schwartz, Thomas M. Jessell, Steven A. Siegelbaum, and A. J. Hudspeth (New York: McGraw-Hill, 2013), 654 – 681.

4 A. J. Hudspeth, Frank Jülicher, and Pascal Martin, "A Critique of the Critical Cochlea: Hopf – a Bifurcation – is Better than None," *Journal of Neurophysiology* 104 (2010): 1219 – 1229.

5 A. J. Hudspeth, "Integrating the Active Process of Hair Cells with Cochlear Function," *Nature Reviews Neuroscience* 15 (2014): 600 – 614.

6 Dylan Thomas, *The Poems of Dylan Thomas*, ed. Daniel Jones (New York: New Directions, 1971), 195 – 196.

7 Daniel P. Walker, *Studies in Musical Science in the Late Renaissance* (London: London University Press, 1978), 123 – 170.

Remembering

Larry R. Squire & John T. Wixted

Abstract: A major development in understanding the structure and organization of memory was the identification of the medial temporal lobe memory system as one of the brain systems that support memory. Work on this topic began in the 1950s with the study of the noted amnesic patient H.M. and culminated in studies of an animal model of human memory impairment in the nonhuman primate. These discoveries opened new frontiers of research concerned with the functional specialization of structures within the medial temporal lobe, the existence of multiple memory systems, the process of memory consolidation, and the role of neural replay and sleep in the consolidation process. This work also led to new insights about how and where memories are ultimately stored in the brain. All of this research has improved our understanding of how memory is affected by normal aging and why it is so profoundly impaired by the pathological processes associated with dementia.

LARRY R. SQUIRE, a Fellow of the American Academy since 1995, is Research Career Scientist at the Veterans Affairs San Diego Healthcare System and Distinguished Professor of Psychiatry, Neurosciences, and Psychology at the University of California, San Diego.

JOHN T. WIXTED is Distinguished Professor of Psychology at the University of California, San Diego.

*(*See endnotes for complete contributor biographies.)*

Memory is a large topic, growing out of the fundamental fact that the experiences we have can modify the nervous system such that our mental life and our behavior can be different than they were in the past. The study of memory ranges widely – from cellular and molecular questions about the nature of synaptic change to questions about what memory is, whether it is one thing or many, which brain systems support memory, and how those systems operate. We will consider in particular the structure and organization of memory with a focus on brain systems.

The idea that functions of the nervous system can be localized was well accepted by the end of the nineteenth century. Yet these ideas concerned mainly sensory-motor functions and language and did not speak to the topic of memory itself. In the early twentieth century, an influential program of research in the rat concluded that memory is not localized but is distributed through the neocortex (the outer layer of the cerebral hemispheres of the brain of mammals in-

doi:10.1162/DAED_a_00317

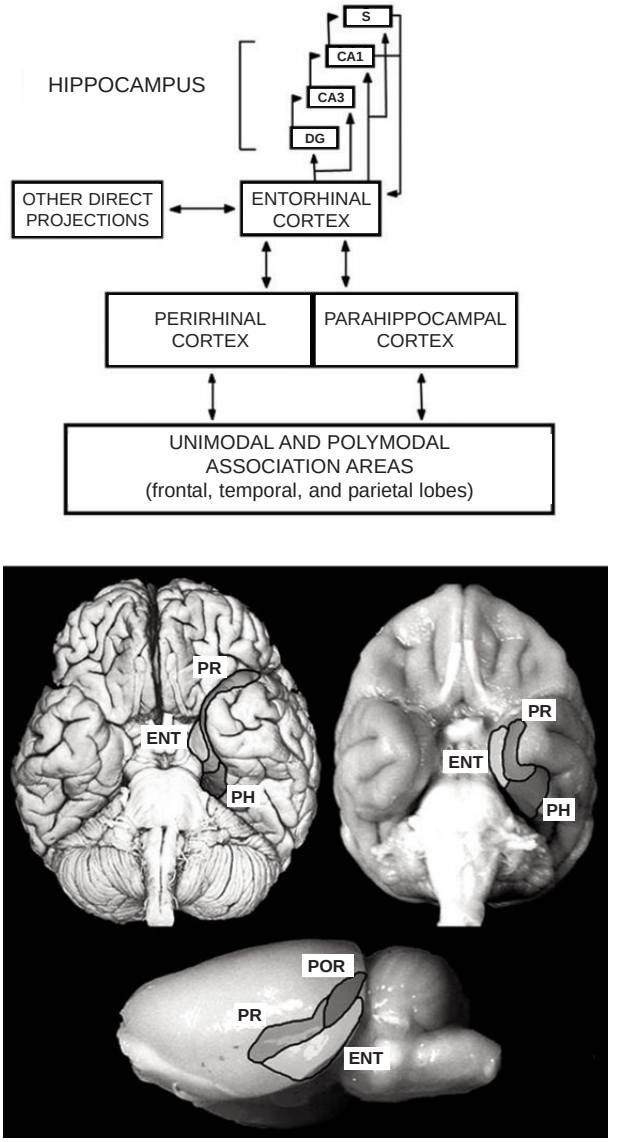
volved in higher functions such as sensory perception, attention, memory, and action), such that each region contributes equivalently to the whole.[1] Memory was thought to be distributed and well integrated with intellectual and perceptual functions, and no particular brain region was thought to be dedicated to memory function.

All of this changed in the 1950s when profound effects on memory were reported following a bilateral medial temporal lobe resection (the removal of the inner structures of the temporal lobe) carried out in the patient known as H.M.[2] This experimental surgery successfully relieved H.M.'s severe epilepsy, as was intended, but it also resulted in severe and debilitating forgetfulness, which occurred against a background of apparently intact intellectual and perceptual functions. For example, the patient could copy a complex drawing as well as controls, suggesting that his ability to perceive visual information was intact; and he could continuously rehearse (and then repeat back) a string of five or six digits as well as controls, suggesting that his "working memory" was also intact. But when his attention was diverted, he soon forgot the drawing and the digits. Early descriptions of H.M. can be said to have inaugurated the modern era of memory research and strongly influenced the direction of subsequent work. Most significantly, this work identified for the first time a particular area of the brain as important for memory.

H.M.'s bilateral lesion included the hippocampus, amygdala, and the adjacent parahippocampal gyrus. The immediate question was which structures within this large surgical removal were responsible for his circumscribed memory impairment; that is, which structures and connections within the human temporal lobe have dedicated memory functions. These matters became understood gradually during the 1980s following the successful development of an animal model of human amnesia in the nonhuman primate.[3] The important structures proved to be the hippocampus and the adjacent entorhinal, perirhinal, and parahippocampal cortices, which make up much of the parahippocampal gyrus (Figure 1).[4] (Anatomically related structures in the thalamus and hypothalamus in the diencephalic midline, an area not part of H.M.'s lesion, are also important for memory, but these will not be discussed.) Damage limited to the hippocampus itself causes moderately severe memory impairment, but the impairment is greatly exacerbated when the damage extends to and includes the parahippocampal gyrus (as was the case with H.M.).[5] In all cases, the disorder is characterized most prominently by an impaired ability to form new memories (anterograde amnesia), but also by difficulty in accessing some memories acquired before the onset of the impairment (retrograde amnesia). Memories acquired shortly before the occurrence of a brain lesion (such as during the previous year) tend to be more impaired than memories acquired in the distant past. Thus, the structures that compose the medial temporal lobe memory system are essential for the initial formation of enduring long-term memories as well as for their maintenance and retrieval for a time after learning. The fact that very remote memory tends to be preserved after medial temporal lobe damage indicates that these structures are not the ultimate repository of long-term memory.

Once the important structures of the medial temporal lobe were identified, the question naturally arose whether the different structures have specialized roles. An early view held that the hippocampus plays an especially important role in spatial memory.[6] This idea was based on the common finding that rodents with selective

Figure 1
The Medial Temporal Lobe Memory System



Top: Schematic view of the memory system, which is composed of the hippocampus and the perirhinal, entorhinal, and parahippocampal cortices. In addition to the connections shown here, there are also weak projections from the perirhinal and parahippocampal cortices to the CA1-subiculum border. *Bottom*: Ventral view of a human brain (upper left) and a monkey brain (upper right) and a lateral view of a rat brain (lower center). The major cortical components of the medial temporal lobe are highlighted and outlined. The hippocampus is not visible from the surface and, in the human, lies beneath the structures of the medial temporal lobe. Its anterior extent lies below the posterior entorhinal and perirhinal cortices, and the main body of the hippocampus lies beneath the parahippocampal cortex. In the rat, the parahippocampal cortex is termed the postrhinal cortex. Abbreviations: DG, dentate gyrus; ENT, entorhinal cortex; PH, parahippocampal cortex; POR, postrhinal cortex; PR, perirhinal cortex; S, subicular complex. Source: Adapted from Figure 2 in Larry R. Squire and John T. Wixted, "The Cognitive Neuroscience of Memory since H.M.," *Annual Review of Neuroscience* 34 (2011): 259 – 288.

hippocampal lesions are severely impaired on spatial learning tasks, such as learning to navigate a maze. However, subsequent work involving humans and monkeys with selective hippocampal lesions demonstrated pronounced spatial and nonspatial memory impairment. For example, patients with hippocampal lesions were impaired in their ability to recognize words that had appeared on an earlier list – a task with no obvious spatial component.[7] Findings like these suggest that the hippocampus plays a broader role in memory encoding and consolidation (the gradual process by which a temporary, labile memory is transformed into a more stable, long-lasting form).

Another popular idea about specialization of function within the medial temporal lobe was based on a long-standing psychological distinction between familiarity and recollection.[8] Familiarity involves knowing only that an item has been previously encountered (for example, when you recognize a face but cannot recall who the person is), and recollection involves recalling specific details about the prior encounter (such as recalling where and when you met the familiar person). Initially, a number of findings were interpreted to mean that hippocampal lesions selectively impair the recollection process but leave memory based on familiarity intact.[9] In addition, neuroimaging studies were often interpreted to mean that recollection-based decisions generate elevated activity in the hippocampus, whereas familiarity-based decisions generate elevated activity in other medial temporal lobe structures, particularly the perirhinal cortex.[10] However, subsequent studies found that bilateral hippocampal lesions in humans have comparable effects on recollection and familiarity, and neuroimaging studies found that both familiarity-based and recollection-based recognition generate elevated hippocampal activity when both kinds of memory are strong.[11] Thus, the specialization of function within the medial temporal lobe does not seem to be informed by this distinction.

Because the functions of the different medial temporal lobe structures do not apparently divide up along the lines of spatial versus nonspatial memory or recollection versus familiarity, we must look elsewhere to identify functional differences between the structures. An important consideration is the fact that the inputs to each structure are quite different.[12] For example, the perirhinal cortex receives the majority of its cortical input from areas supporting visual object perception. Thus, the perirhinal cortex may be particularly important for forming memories of visual objects. Similarly, the parahippocampal cortex receives significant input from areas supporting spatial processing (for example, the ability to perceive that objects A and B are closer together than objects C and D). This area may therefore be particularly important for forming memories about the spatial locations of objects. A growing body of evidence is consistent with these ideas.[13] That is, the functional specialization of different medial temporal lobe structures is sensibly related to the domain of information they process – information that is carried to these structures from upstream regions supporting different kinds of perceptual processing.[14]

Within the medial temporal lobe, the hippocampus is the ultimate recipient of convergent projections from the entorhinal, perirhinal, and parahippocampal cortices. Thus, the hippocampus itself is in a position to play a role in the encoding and consolidation of all aspects of an experience (its visual, spatial, auditory, and olfactory qualities, as well as other contextual information). These anatomical facts can therefore explain why damage to the hippocampus results in broad memory impairment that covers all modalities and ex-

tends across multiple domains. Current studies are using new genetic methods in mice and other techniques to analyze the separate contributions of specific connections and cell types within the hippocampus.[15]

The memory impairment associated with medial temporal lobe lesions is narrower than once thought, because not all forms of learning and memory are affected. The first clue came in 1962 when H.M. was found capable of acquiring a motor skill (mirror drawing) over a period of three days, though he could not recall these periods of practice. While this finding showed that memory is not unitary, discussions at the time tended to set aside motor skills as a special case representing a less cognitive form of memory. The suggestion was that the rest of memory is of one piece and is dependent on medial temporal lobe structures.

Yet during the subsequent years, it was discovered that motor-skill learning is but one example of a large domain of abilities that are independent of the medial temporal lobe. An early discovery was that perceptual and cognitive skills – not just motor skills – are intact in patients like H.M. Thus, memory-impaired patients acquired at a normal rate the skill of reading mirror-reversed words, despite poor memory for the words themselves.[16] This finding led to the proposal of a brain-based distinction between declarative and procedural knowledge. Declarative knowledge referred to knowledge available as conscious recollections about facts and events. Procedural knowledge referred to skill-based information: knowledge expressed through performance rather than recollection.

Soon after this discovery was made, the phenomenon of priming was also found to be spared in amnesia.[17] Priming refers to an improved ability to detect or identify stimuli based on a recent encounter with the same or related stimuli. For example, memory-impaired patients could (like healthy volunteers) name recently presented object drawings one hundred milliseconds faster than new drawings, despite having poor memory for the drawings themselves.[18] Perhaps the most compelling evidence for the independence of priming and ordinary memory ability was that severely amnesic patients can exhibit fully intact priming for words while performing only at chance levels on conventional recognition memory tests for the same words.[19]

Another important insight was the idea that the neostriatum (a subcortical region of the brain that includes the caudate nucleus and putamen), and not the medial temporal lobe, is important for the sort of gradual, feedback-guided learning that results in habit memory.[20] For example, memory-impaired patients learned tasks at a normal rate when the outcome of each learning trial was determined probabilistically, and performance therefore needed to be based on a gut feeling rather than on conscious memory of past events.[21] Work with experimental animals was also the source of new insights, including the discovery in the early 1980s that the cerebellum is essential for delay eyeblink conditioning,[22] a kind of learning entirely preserved after hippocampal lesions.[23] Still other types of learning, which involve attaching a positive or negative valence to a stimulus (as in fear conditioning), depend on the amygdala.[24]

Given the variety of tasks explored in these studies and the number of brain structures implicated, an account of memory based on a two-part dichotomy (declarative versus procedural) began to seem too simplistic. Accordingly, the perspective eventually shifted to a framework that accommodated more than two memory systems. At that time, the umbrella term "nondeclarative memory" was introduced with the intention of distinguishing between

declarative memory (which refers to one memory system) and other types of memory (in which several additional systems are involved).[25] Figure 2 illustrates this idea.[26]

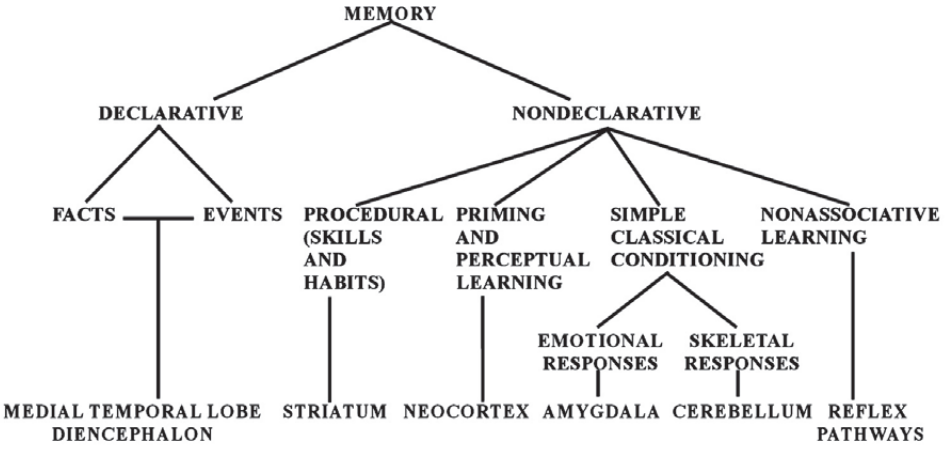
Declarative memory is what the term *memory* signifies when we use it in everyday language. The stored representations are flexible and thought to be accessible to conscious awareness. Declarative memory is representational; it provides a way to model the external world and is either true or false. In contrast, nondeclarative memory is neither true nor false: it is dispositional and occurs as modifications within specialized performance systems. Thus, the various memory systems can be distinguished in terms of the different kinds of information they process and the principles by which they operate. These systems work in parallel to support behavior. For example, an aversive event in childhood (such as being knocked down by a large dog) can lead to an enduring declarative memory of the event itself (dependent on the hippocampus and related structures) as well as a long-lasting, nondeclarative fear of dogs (a phobia, dependent on the amygdala) that is experienced as part of the personality rather than as a memory.

The hippocampus and related structures in the medial temporal lobe have a time-limited role in the formation and storage of memory. Two lines of work underlie this idea. First, damage to these structures typically spares remote memory and impairs more recent memory in a temporally graded fashion. In humans, hippocampal lesions affect memory for up to a few years after learning. In experimental animals (usually rats or mice), similar damage impairs memory for up to thirty days after learning.[27] Thus, long-term, stable memory develops more slowly in humans than in experimental animals. Discussion in the field continues about the possible special status of spatial memory and autobiographical memory in humans and the idea that these forms of memory might depend on medial temporal lobe structures as long as memory persists.[28] Yet there are reports of patients with medial temporal lobe lesions in whom remote spatial and autobiographical memory has been spared.[29]

The second line of work involves studies of experimental animals that track neural activity or structural changes in the hippocampus and neocortex after learning. For example, expression patterns of activity-related genes like c-Fos describe gradually decreasing activity in the hippocampus after learning and parallel increases in activity in a number of cortical regions.[30] These findings and others describe the increasing importance of distributed cortical regions for the representation of memory as time passes after learning.[31] Similar findings have been obtained in neuroimaging studies; for example, when volunteers attempt to recall news events that occurred anywhere from one to thirty years earlier.[32] The idea is not that memory is literally transferred from the hippocampus to the neocortex. Memory is always in the neocortex, but gradual changes occur to increase the complexity, distribution, and connectivity of memory representations among multiple cortical regions. At the same time the role of the hippocampus gradually diminishes (Figure 3).

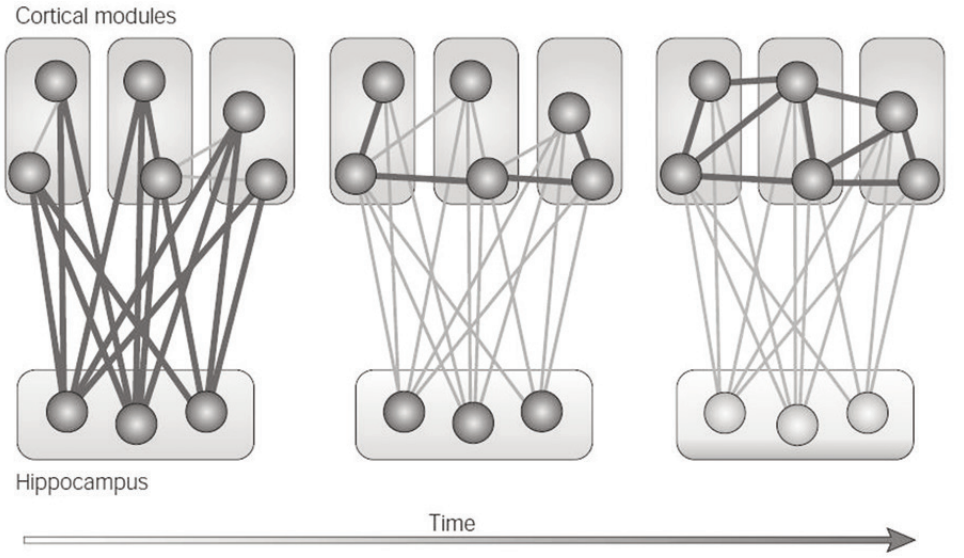
One way to view this process is to suppose that a time-and-place-specific new memory (a so-called episodic memory) is represented initially by an ensemble of distributed changes in the neocortex and by changes in the hippocampus (and anatomically related structures) as well. The neocortical ensemble is viable so long as the episode is maintained within active memory. However, when one's attention is directed elsewhere, a problem arises. How can the unique distribution of sites that represent this new memory be revivified by

Figure 2
Organization of Mammalian Long-Term Memory Systems



The figure lists the brain structures thought to be especially important for each form of declarative and nondeclarative memory. In addition to its central role in emotional learning, the amygdala is able to modulate the strength of both declarative and nondeclarative memory. Source: Figure prepared by Larry R. Squire.

Figure 3
Consolidation of Memory in the Neocortex



Encoding of new information initially engages the hippocampus and a distributed set of specialized cortical areas (left panel). Subsequent reactivation of this hippocampal-cortical network progressively strengthens cortico-cortical connections or establishes new ones (middle panel). Eventually the cortico-cortical connections are sufficiently strong and stable for memory to be maintained and retrieved independently of the hippocampus (right panel). Source: Paul W. Frankland and Bruno Bontempi, "The Organization of Recent and Remote Memories," *Nature Reviews Neuroscience* 6 (2005): 119 – 130.

Remembering

unaided recall or after the presentation of a partial reminder? The notion is that remembering becomes possible because medial temporal lobe structures, by way of their widespread, divergent connections to the neocortex, effectively bind together the distributed neocortical sites that together constitute the new memory. This connectivity supports the capacity for remembering during the consolidation process until the connectivity among the relevant cortical sites becomes strong enough to represent a stable memory without the support of the medial temporal lobe.

A long-standing idea, which has received renewed attention in recent years, is that retrieval of memory provides an opportunity for updating or modulating what was originally learned and even the possibility of severely disrupting it.[33] The process by which a long-term memory transiently returns to a labile state (and then re-stabilizes) has been termed reconsolidation. Although it is clear that memory can be modified or distorted by memory retrieval, questions remain about the conditions under which memory can actually be abolished. Some studies in experimental animals report that a reactivated memory can be impaired but that the disruption is transient.[34] Other studies in animals report that only recent memories (ones that are one or seven days old, but not fourteen or twenty-eight days old) can be impaired after reactivation.[35]

Consolidation presumably requires some relatively long-lasting form of communication between the medial temporal lobe and the neocortex. One proposal for how this could be accomplished is through the phenomenon of neural replay. Recordings of neural activity in rodents showed that firing sequences of hippocampal neurons during waking behavior are then spontaneously replayed during subsequent slow-wave sleep.[36] Later it was found that hippocampal replay was coordinated with firing patterns in the visual cortex, which is consistent with the idea that a dialogue occurs between hippocampus and neocortex.[37] This coordination could be part of the process by which recent memories eventually become consolidated remote memories. Interestingly, disrupting replay activity in rodents during a rest period (filled by quiet wakefulness and slow-wave sleep) following spatial learning impairs later memory for the task.[38]

These studies with rodents led to conceptually similar studies with humans. For example, volunteers memorized the locations of card pairs on a computer screen while being exposed to a particular odor (the smell of a rose). Later, odor reexposure, specifically during slow-wave sleep, increased hippocampal activity (measured by neuroimaging) and lessened forgetting of the card pair locations following sleep.[39] In another study, the hippocampus and parahippocampal gyrus were active while participants learned routes in a virtual reality environment and were active again during subsequent slow-wave sleep.[40] The degree of activation during slow-wave sleep correlated with memory performance the next day. Studies like these have been interpreted to mean that consolidation results from the reactivation of newly encoded hippocampal representations, specifically during slow-wave sleep.[41]

An important question is whether neural replay and the consolidation process are specific to slow-wave sleep or whether these events might occur whenever the brain is not actively encoding new memories, such as during quiet wakefulness.[42] In rodents, neural replay can occur during wakefulness.[43] Moreover, in a neuroimaging study with humans, coordinated hippocampal-cortical activity occurred during a rest period that followed learning, and this activity predicted later memory performance.[44] Accordingly, an intriguing pos-

sibility is that the neural replay activity proposed to underlie memory consolidation may occur whenever the brain is in a quiet state (not just during slow-wave sleep).

Where are memories ultimately stored in the brain? A variety of evidence has converged on the view that the different aspects of remembered information are stored in the same regions of the brain that initially perform the processing and analysis of that information. According to this view, remembering a previous experience consists of the coordinated reactivation of the distributed neocortical regions that were activated during initial perceptual processing.[45] While the memory is still new, this reactivation of distributed cortical activity depends on the hippocampus and other medial temporal lobe structures, but once memory is fully consolidated, reactivation can occur within the neocortex itself. Each neocortical region operates within a specific domain and stores only the features of an experience – such as visual, auditory, or spatial information – that belong to that domain. Thus, as proposed by psychologist Karl Lashley long ago, memories are distributed throughout the neocortex.[46] However, contrary to his view, memory is not uniformly distributed. Some areas are more important for storing the visual aspects of an experience, and other areas are more important for storing other aspects.

An implication of this view is that neocortical lesions that selectively impair perceptual processing in a particular domain (such as the perceptual processing of color) should also cause correspondingly specific anterograde and retrograde memory impairment within the same domain. This circumstance is illustrated by "The Case of the Colorblind Painter," a case described by the neurologist Oliver Sacks.[47] An accomplished painter was involved in an automobile accident at the age of sixty-five, which rendered him color-blind. The disability was striking: he could discriminate between wavelengths of light, even though the different wavelengths gave rise to the perception of various shades of gray rather than the perception of different colors. Because his condition was acquired (it was not congenital), it was possible to interrogate not only his ability to form new color memories, but also the status of previously established memories that had once included the subjective experience of color. The case description leaves little doubt that the patient's experience – both going forward and looking back – was now completely (and selectively) devoid of color. Although he retained abstract semantic knowledge of color, he could neither perceive nor later remember the color of objects presented to him (anterograde impairment). In addition, he could not subjectively experience color in his earlier (and once chromatic) memories. For example, he knew that his lawn was green, but he reported that he could no longer visualize it in green when he tried to remember what it once looked like.

Note the difference between the effect of this cortical lesion on memory and the effect of bilateral medial temporal lobe lesions. With respect to remote memories that have already been fully consolidated, medial temporal lobe lesions have little effect. In contrast, focal cortical lesions can selectively abolish one feature (like color) of a long-consolidated memory. With respect to new experiences, bilateral medial temporal lesions lead to severe anterograde amnesia (no subsequent memory for a recent experience). In contrast, focal cortical lesions of the kind suffered by the painter prevent the encoding and retrieval of only one aspect of the experience (color in his case). Because the processing of color in the painter's neocortex was impaired, his experience of color was eliminated in both perception and memory.

Remembering

Selective deficits in long-term knowledge of the kind suffered by the painter are not limited to perceptual experience. Semantic knowledge (knowledge about objects, facts, and word meanings) is also stored in neocortical regions that can be selectively damaged.[48] Thus, damage limited to lateral regions of patients' temporal lobe (close to, but not including, medial temporal lobe structures) can disrupt previously stored information – such as what an animal looks or sounds like. Such patients have difficulty naming pictures of animals and providing information about them. Other patients with damage to the parietal cortex can have difficulty identifying small manipulable objects (like spoons and brushes) and knowing how to use them. Neuroimaging studies support the findings from lesion studies and show that the properties of objects, together with how they are perceived and used, influence which brain areas store long-term knowledge about their identity.[49]

The information in the preceding sections helps illuminate some of the memory deficits associated with normal aging and dementia. One of the most common experiences associated with normal aging is the decline in memory function. Oftentimes, the memory difficulty is characterized as poor "short-term" memory. In its common usage, a short-term memory problem means having trouble remembering recent experiences (such as when someone tells a story for the second time without remembering having told it before) while at the same time having no trouble remembering events from decades ago. Older adults who exhibit these symptoms are having difficulty encoding and consolidating new memories, while memories that were acquired and consolidated long ago are easy to retrieve. These changes in memory ability are related to changes within medial temporal lobe structures. In experimental animals, the dentate gyrus within the hippocampus is most sensitive to the effects of aging.[50] Studies in humans have reported between 1 and 2 percent annual hippocampal atrophy in non-demented adults older than fifty-five years.[51] Aerobic exercise can reverse age-related volume loss by one to two years.[52]

Alzheimer's disease, the most common form of dementia, is a progressive neurodegenerative condition. It is a distinct condition, not an acceleration of the normal aging process. The first targets of the disease are the entorhinal cortex and the CA1 field of the hippocampus, which explains why memory is especially affected in its early stages.[53] The rate of hippocampal volume loss is at least 2.5 times greater in Alzheimer's disease than in normal aging.[54] The disease progresses to involve intellectual functions quite broadly. The neocortex becomes involved (though sensory and motor areas are relatively spared) and patients develop difficulty with language, problem solving, calculation, and judgment.

Semantic dementia, another progressive disorder, begins elsewhere in the brain and is associated with a different pattern of symptoms.[55] This condition prominently involves atrophy of the anterior and lateral temporal lobes.[56] Unlike patients with Alzheimer's disease, these patients have severe loss of previously stored and long-consolidated semantic knowledge (that is, loss of conceptual knowledge about objects, facts, and word meanings). Yet their ability to form new memories can be relatively spared. Thus, patients could recognize which drawings of animals they had seen recently but failed at tests of conceptual knowledge about the same items.[57] Not just the name of the item is lost – the concept itself is degraded.

The understanding of memory has changed in ways that might have seemed

revolutionary to Karl Lashley when he searched for sites of memory storage in the brains of rats.[58] All that has been learned about the structure and organization of memory and about brain systems is the result of basic, fundamental research, mostly in rodents, monkeys, and humans. Although we did not review it here, much has also been learned from studies of the cellular and molecular basis of memory, an enterprise that has depended heavily on mice as well as invertebrate animals like *Aplysia* and *Drosophila*. As this work continues, one can expect not only new insights into how memory operates but also improved understanding of human health and disease, including improved ways to diagnose, treat, and prevent the diseases that affect memory.

ENDNOTES

* Contributor Biographies: LARRY R. SQUIRE, a Fellow of the American Academy since 1996, is Research Career Scientist at the Veterans Affairs San Diego Healthcare System and Distinguished Professor of Psychiatry, Neurosciences, and Psychology at the University of California, San Diego. He is the author of *Memory and Brain* (1987) and *Memory: From Mind to Molecules* (with Eric Kandel, 2009), and is also the Senior Editor of *Fundamental Neuroscience* (fourth edition, 2013).

JOHN T. WIXTED is Distinguished Professor of Psychology at the University of California, San Diego. He is an Associate Editor of *Psychological Review*, and his work has been published in journals such as *Trends in Neurosciences*, *Neuropsychologia*, and *Psychological Review*; and in volumes such as *The Oxford Handbook of Cognitive Neuroscience* (edited by Stephen Kosslyn and Kevin N. Ochsner, 2013) and *Memory and Law* (edited by Lynn Nadel and Walter Sinnott-Armstrong, 2012).

Authors' Note: The preparation of this article was supported by the Medical Research Service of the Department of Veterans Affairs and NIMH Grant 24600.

1 Karl S. Lashley, *Brain Mechanisms and Intelligence: A Quantitative Study of Injuries to the Brain* (Chicago: University of Chicago Press, 1929).

2 William B. Scoville and Brenda Milner, "Loss of Recent Memory after Bilateral Hippocampal Lesions," *Journal of Neurology, Neurosurgery, and Psychiatry* 20 (1957): 11 – 21.

3 Mortimer Mishkin, "Memory in Monkeys Severely Impaired by Combined but Not by Separate Removal of Amygdala and Hippocampus," *Nature* 273 (1978): 297 – 298.

4 Larry R. Squire and Stuart Zola-Morgan, "The Medial Temporal Lobe Memory System," *Science* 253 (1991): 1380 – 1386.

5 S. Zola-Morgan, L. R. Squire, and D. G. Amaral, "Human Amnesia and the Medial Temporal Region: Enduring Memory Impairment Following a Bilateral Lesion Limited to Field CA1 of the Hippocampus," *The Journal of Neuroscience* 6 (1986): 2950 – 2967; and Nancy L. Rempel-Clower et al., "Three Cases of Enduring Memory Impairment Following Bilateral Damage Limited to the Hippocampal Formation," *The Journal of Neuroscience* 16 (1996): 5233 – 5255.

6 John O'Keefe and Lynn Nadel, *The Hippocampus as a Cognitive Map* (London: Oxford University Press, 1978).

7 Jonathan M. Reed and Larry R. Squire, "Impaired Recognition Memory in Patients with Lesions Limited to the Hippocampal Formation," *Behavioral Neuroscience* 111 (1997): 667 – 675.

8 Richard C. Atkinson and Jim F. Juola, *Contemporary Developments in Mathematical Psychology*, ed. David H. Krantz, Richard C. Atkinson, and Patrick Suppes (San Francisco: W. H. Freeman, 1974), 243 – 290; and G. Mandler, "Recognizing: The Judgment of Previous Occurrence," *Psychological Review* 87 (1980): 252 – 271.

[9] Malcolm W. Brown and John P. Aggleton, "Recognition Memory: What Are the Roles of the Perirhinal Cortex and Hippocampus?" *Nature Reviews Neuroscience* 2 (2001): 51–61.

[10] H. Eichenbaum, A. P. Yonelinas, and C. Ranganath, "The Medial Temporal Lobe and Recognition Memory," *Annual Review of Neuroscience* 30 (2007): 123–152.

[11] Z. Song et al., "Impaired Capacity for Familiarity after Hippocampal Damage," *Proceedings of the National Academy of Sciences* 108 (2011): 9655–9660; Christine N. Smith, John T. Wixted, and Larry R. Squire, "The Hippocampus Supports Both Recollection and Familiarity When Memories Are Strong," *The Journal of Neuroscience* 31 (2011): 15693–17502; and Z. Song, A. Jeneson, and L. R. Squire, "Medial Temporal Lobe Function and Recognition: A Novel Approach to Separating the Contribution of Recollection and Familiarity," *The Journal of Neuroscience* 31 (2011): 16026–16032.

[12] W. A. Suzuki and D. G. Amaral, "Topographic Organization of the Reciprocal Connections between the Monkey Entorhinal Cortex and the Perirhinal and Perihippocampal Cortices," *The Journal of Neuroscience* 14 (1994): 1856–1877; and L. R. Squire, "Mechanisms of Memory," *Science* 232 (1986): 1612–1619.

[13] Elizabeth A. Buffalo, Patrick S. F. Bellgowan, and Alex Martin, "Distinct Roles for Medial Temporal Lobe Structures in Memory for Objects and Their Locations," *Learning & Memory* 13 (2006): 638–643; Bernhard P. Staresina, Katherine D. Duncan, and Lila Davachi, "Perirhinal and Parahippocampal Cortices Differentially Contribute to Later Recollection of Object- and Scene-Related Event Details," *The Journal of Neuroscience* 31 (2011): 8739–8747; Jackson C. Liang, Anthony D. Wagner, and Alison R. Preston, "Content Representation in the Human Medial Temporal Lobe," *Cerebral Cortex* 23 (2013): 80–96; and Bernhard P. Staresina, Elisa Cooper, and Richard N. Henson, "Reversible Information Flow across the Medial Temporal Lobe: The Hippocampus Links Cortical Modules During Memory Retrieval," *The Journal of Neuroscience* 33 (2013): 14184–14192.

[14] John T. Wixted and Larry R. Squire, "The Medial Temporal Lobe and the Attributes of Memory," *Trends in Cognitive Science* 15 (2011): 210–217.

[15] Michael A. Yassa and Craig E. L. Stark, "Pattern Separation in the Hippocampus," *Trends in Neurosciences* 34 (2011): 515–525; and Xu Liu et al., "Optogenetic Stimulation of a Hippocampal Engram Activates Fear Memory Recall," *Nature* 1038 (2012): 381–385.

[16] Neal J. Cohen and Larry R. Squire, "Preserved Learning and Retention of Pattern Analyzing Skill in Amnesia: Dissociation of Knowing How and Knowing That," *Science* 210 (1980): 207–209.

[17] Endel Tulving and Daniel L. Schacter, "Priming and Human Memory Systems," *Science* 247 (1990): 301–306; Elizabeth K. Warrington and Rosaleen A. McCarthy, "Categories of Knowledge: Further Fractionations and an Attempted Integration," *Brain* 110 (1987): 1273–1296; and Daniel L. Schacter and Randy L. Buckner, "Priming and the Brain," *Neuron* 20 (1998): 185–195.

[18] Carolyn Backer Cave and Larry R. Squire, "Intact and Long-Lasting Repetition Priming in Amnesia," *The Journal of Experimental Psychology: Learning, Memory and Cognition* 18 (1992): 509–520.

[19] Stephan B. Hamann and Larry R. Squire, "Intact Perceptual Memory in the Absence of Conscious Memory," *Behavioral Neuroscience* 111 (1997): 850–854.

[20] Mortimer Mishkin et al., "Memories and Habits: Two Neural Systems," in *Neurobiology of Human Learning and Memory*, ed. G. Lynch, J. L. McGaugh, and W. M. Weinberger (New York: Guilford, 1984), 65–77.

[21] Barbara J. Knowlton, Jennifer A. Mangels, and Larry R. Squire, "A Neostriatal Habit Learning System in Humans," *Science* 273 (1996): 1399–1402.

[22] Delay eyeblink conditioning is a form of Pavlovian conditioning in which a conditioned stimulus (such as a tone) is presented and remains on until the unconditioned stimulus (such as a puff of air to the eye) is presented. The two stimuli overlap and co-terminate. D. A. McCormick

et al., "Initial Localization of the Memory Trace for a Basic Form of Learning," *Proceedings of the National Academy of Sciences* 79 (1982): 2731 – 2735.

23 Robert E. Clark and Larry R. Squire, "Awareness and the Conditioned Eyeblink Response," in *Eyeblink Classical Conditioning, Vol. 1: Applications in Humans*, ed. Diana S. Woodruff-Pak and Joseph E. Steinmetz (Norwell, Mass.: Kluwer Academic Publishers, 2000), 229 – 251; and K. M. Christian and R. F. Thompson, "Neural Substrates of Eyeblink Conditioning: Acquisition and Retention," *Learning & Memory* 10 (2003): 427 – 455.

24 Joseph LeDoux, *The Emotional Brain* (New York: Simon & Schuster, 1996).

25 Larry R. Squire and Stuart Zola-Morgan, "Memory: Brain Systems and Behavior," *Trends in Neurosciences* 11 (1988): 170 – 175.

26 For earlier versions of this diagram, see Squire, "Mechanisms of Memory."

27 Larry R. Squire and Peter J. Bayley, "The Neuroscience of Remote Memory," *Current Opinion in Neurobiology* 17 (2007): 185 – 196.

28 Morris Moscovitch et al., "The Cognitive Neuroscience of Remote Episodic, Semantic and Spatial Memory," *Current Opinion in Neurobiology* 16 (2006): 179 – 190.

29 Squire et al., "The Neuroscience of Remote Memory."

30 Paul W. Frankland and Bruno Bontempi, "The Organization of Recent and Remote Memories," *Nature Reviews Neuroscience* 6 (2005): 119 – 130.

31 Leonardo Restivo et al., "The Formation of Recent and Remote Memory is Associated with Time-Dependent Formation of Dendritic Spines in the Hippocampus and Anterior Cingulate Cortex," *The Journal of Neuroscience* 29 (2009): 8206 – 8214.

32 C. N. Smith and L. R. Squire, "Medial Temporal Lobe Activity During Retrieval of Semantic Memory Is Related to the Age of the Memory," *The Journal of Neuroscience* 29 (2009): 930 – 938.

33 Elizabeth F. Loftus, "Planting Misinformation in the Human Mind: A 30-Year Investigation of the Malleability of Memory," *Learning & Memory* 12 (2005): 361 – 366; Jonathan L. Lee, "Reconsolidation: Maintaining Memory Relevance," *Trends in Neuroscience* 32 (2009): 413 – 420; Yadin Dudai, "The Restless Engram: Consolidations Never End," *The Annual Review of Neuroscience* 35 (2012): 227 – 247; Peggy L. St. Jacques, Christopher Olm, and Daniel L. Schacter, "Neural Mechanisms of Reactivation-Induced Updating that Enhance and Distort Memory," *Proceedings of the National Academy of Sciences* 110 (49): 19671 – 19678; and Karim Nader, Glenn E. Schafe, and Joseph E. LeDoux, "Fear Memories Require Protein Synthesis in the Amygdala for Reconsolidation after Retrieval," *Nature* 406 (2000): 722 – 726.

34 K. Matthew Lattal and Ted Abel, "Behavioral Impairments Caused by Injections of the Protein Synthesis Inhibitor Anisomycin after Contextual Retrieval Reverse with Time," *Proceedings of the National Academy of Sciences* 101 (2004): 4667 – 4672; and Ann E. Power et al., "Anisomycin Infused into the Hippocampus Fails to Block 'Reconsolidation' but Impairs Extinction: The Role of Re-Exposure Duration," *Learning & Memory* 13 (2006): 27 – 34.

35 Maria H. Milekic and Cristina M. Alberini, "Temporally Graded Requirement for Protein Synthesis Following Memory Reactivation," *Neuron* 36 (2002): 521 – 525.

36 M. A. Wilson and B. L. McNaughton, "Reactivation of Hippocampal Ensemble Memories During Sleep," *Science* 265 (1994): 676 – 679.

37 Daoyun Ji and Matthew A. Wilson, "Coordinated Memory Replay in the Visual Cortex and Hippocampus During Sleep," *Nature Neuroscience* 10 (2007): 100 – 107.

38 Valérie Ego-Stengel and Matthew A. Wilson, "Disruption of Ripple-Associated Hippocampal Activity During Rest Impairs Spatial Learning in the Rat," *Hippocampus* 20 (2010): 1 – 10.

39 Björn Rasch et al., "Odor Cues During Slow-Wave Sleep Prompt Declarative Memory Consolidation," *Science* 315 (2007): 1426 – 1429.

40 Philippe Peigneux et al., "Are Spatial Memories Strengthened in the Human Hippocampus During Slow-Wave Sleep?" *Neuron* 44 (2004): 535 – 545.

41 Marion Inostroza and Jan Born, "Sleep for Preserving and Transforming Episodic Memory," *Annual Review of Neuroscience* 36 (2013): 79 – 102.

42 Sara C. Mednick et al., "An Opportunistic Theory of Cellular and Systems Consolidation," *Trends in Neurosciences* 34 (2011): 504 – 514.

43 Mattias P. Karlsson and Loren M. Frank, "Awake Replay of Remote Experiences in the Hippocampus," *Nature Neuroscience* 12 (2009): 913 – 918.

44 Arielle Tambini, Nicholas Ketz, and Lila Davachi, "Enhanced Brain Correlations During Rest are Related to Memory for Recent Experiences," *Neuron* 65 (2010): 280 – 290.

45 E. De Renzi, "Memory Disorders Following Focal Neocortical Damage," *Philosophical Transactions of the Royal Society B: Biological Sciences* 298 (1982): 73 – 83; Mortimer Mishkin, "A Memory System in the Monkey," *Philosophical Transactions of the Royal Society B: Biological Sciences* 298 (1982): 85 – 92; Larry R. Squire, *Memory and Brain* (New York: Oxford University Press, 1987); and Antonio R. Damasio, "Time-Locked Multiregional Retroactivation: A Systems-Level Proposal for the Neural Substrates of Recall and Recognition," *Cognition* 33 (1989): 25 – 62.

46 K. S. Lashley, *Brain Mechanisms and Intelligence: A Quantitative Study of Injuries to the Brain* (Chicago: Chicago University Press, 1929).

47 Oliver Sacks, "The Case of the Colorblind Painter," in Oliver Sacks, *An Anthropologist on Mars: Seven Paradoxical Tales* (New York: Random House, 1995), 3 – 41.

48 Warrington and McCarthy, "Categories of Knowledge: Further Fractionations and an Attempted Integration."

49 Alex Martin, "The Representation of Object Concepts in the Brain," *Annual Review of Psychology* 58 (2007): 25 – 45.

50 Scott A. Small et al., "Imaging Correlates of Brain Function in Monkeys and Rats Isolates a Hippocampal Subregion Differentially Vulnerable to Aging," *Proceedings of the National Academy of Sciences* 101 (2004): 7181 – 7186.

51 C. R. Jack, Jr., et al., "Rate of Medial Temporal Lobe Atrophy in Typical Aging and Alzheimer's Disease," *Neurology* 51 (1998): 993 – 999.

52 Kirk I. Erickson et al., "Exercise Training Increases Size of Hippocampus and Improves Memory," *Proceedings of the National Academy of Sciences* 108 (2011): 3017 – 3022.

53 Bradley T. Hyman et al., "Alzheimer's Disease: Cell-Specific Pathology Isolates the Hippocampal Formation," *Science* 225 (1984): 1168 – 1170; and M. J. West, P. D. Coleman, D. G. Flood, and J. C. Troncoso, "Differences in the Pattern of Hippocampal Neuronal Loss in Normal Ageing and Alzheimer's Disease," *Lancet* 17 (1994): 769 – 772.

54 Jack et al., "Rate of Medial Temporal Lobe Atrophy in Typical Aging and Alzheimer's Disease."

55 John R. Hodges and Kim S. Graham, "Episodic Memory: Insights from Semantic Dementia," *Philosophical Transactions of the Royal Society B: Biological Sciences* 1413 (2001): 1423 – 1434.

56 D. A. Levy, P. J. Bayley, and L. R. Squire, "The Anatomy of Semantic Knowledge: Medial vs. Lateral Temporal Lobe," *Proceedings of the National Academy of Sciences* 101 (2004): 6710 – 6715; and Karalyn Patterson, Peter J. Nestor, and Timothy T. Rogers, "Where Do You Know What You Know?" *Nature Reviews Neuroscience* 8 (2007): 976 – 987.

57 K. S. Graham, J. T. Becker, and J. R. Hodges, "On the Relationship between Knowledge and Memory for Pictures: Evidence from the Study of Patients with Semantic Dementia and Alzheimer's Disease," *Journal of the International Neuropsychological Society* 3 (1997): 534 – 544.

58 Lashley, *Brain Mechanisms and Intelligence*.

Sleep, Memory & Brain Rhythms

Brendon O. Watson & György Buzsáki

Abstract: Sleep occupies roughly one-third of our lives, yet the scientific community is still not entirely clear on its purpose or function. Existing data point most strongly to its role in memory and homeostasis: that sleep helps maintain basic brain functioning via a homeostatic mechanism that loosens connections between overworked synapses, and that sleep helps consolidate and re-form important memories. In this review, we will summarize these theories, but also focus on substantial new information regarding the relation of electrical brain rhythms to sleep. In particular, while REM sleep may contribute to the homeostatic weakening of overactive synapses, a prominent and transient oscillatory rhythm called "sharp-wave ripple" seems to allow for consolidation of behaviorally relevant memories across many structures of the brain. We propose that a theory of sleep involving the division of labor between two states of sleep – REM and non-REM, the latter of which has an abundance of ripple electrical activity – might allow for a fusion of the two main sleep theories. This theory then postulates that sleep performs a combination of consolidation and homeostasis that promotes optimal knowledge retention as well as optimal waking brain function.

Sleep is clearly a basic human drive, yet we do not fully understand its purpose or function. One could argue that quiet but conscious rest could be just as efficient as sleep for recuperating certain parts of the body and would be less dangerous, since the brain would not be closed to outside inputs. From the evolutionary point of view, then, unconscious sleep must offer an unseen advantage to the brain.

In attempting to understand the neural implications of sleep and neural activity during sleep, the field has focused on the view – well supported by data – that sleep benefits memory and general neural function. In more recent years this claim has been split into two subdomains: 1) a hypothesis centered on homeostasis, wherein sleep reverses the overelaboration and exhaustion of neural networks brought about by prolonged waking states; and 2) a hypothesis that sleep consolidates important memories for long-term storage. In sleep theory, as in neuroscience, much attention has recently been focused on synaptic connections, which carry information between neurons. Yet at the level of the synapse, these

BRENDON O. WATSON is a clinical psychiatrist and a research fellow at Weill Cornell Medical College at Cornell University and is doing postdoctoral research work at the Buzsáki Lab at the New York University School of Medicine.

GYÖRGY BUZSÁKI is the Biggs Professor of Neural Sciences at the New York University School of Medicine.

*(*See endnotes for complete contributor biographies.)*

doi:10.1162/DAED_a_00318

two theories seem to conflict: while the homeostatic theory states that synapses, in general, are weakened, the consolidation theory states that selected synaptic connections should be strengthened during sleep as a way to consolidate memory.

We seek here to summarize the major concepts in the neuroscience of sleep (and refer the interested reader to a more comprehensive review of the relationship between sleep and memory).[1] We propose that electrical brain rhythms are key physiological features that allow the brain to carry out all aspects of the tasks of sleep and that offer important insight into those tasks. We also seek to determine whether these two apparently opposing views on sleep might be reconciled.

Before proceeding to examine the relationship between sleep and brain rhythms, it is worth reviewing some aspects of brain structure and function that are pertinent to the topic. Our current understanding of the brain is that the basic currency of computation is a collection of electrical signals transferred from one cell to another. This occurs via action potentials (electrical signals within neurons that are triggered after neurons have received sufficient excitatory input) and highly adaptable chemical synaptic contacts (specialized junctions between neurons that allow information to pass between them). The action potential signals are generated by individual neurons at rates ranging from one per minute to tens or even hundreds per second. They are large enough in amplitude to be measured from outside the neuron, and extracellular recordings are often used by neuroscientists as measures of information transmission by a given neuron or population of neurons. The synaptic connections among neurons are relatively sparse and are often structured rather than random, creating functional "circuits"; and perhaps resultantly, volleys of action potentials are often generated by coordinated populations of neurons in a cohesive manner.

All of this complexity must be harnessed and organized somehow. This is partly accomplished through the spatial segregation of neurons into subdivisions of the brain (often referred to as *nuclei* or simply *regions*) such as the hippocampus, the thalamus, or the neocortex. Each region and its interactions are thought to handle specific neural tasks: controlling breathing rhythms, enabling visual perception, handling emotions, or navigating places and memories. To orchestrate these spatially distinct and seemingly task-specific regions, the brain employs a temporal organizational scheme using periodic electrical oscillations (regular fluctuations in electrical potential occurring simultaneously in many neurons, which are measurable even from outside the brain) to achieve two fundamental operations.

The first is perpetual local-global communication, whereby the results of local computations are broadcast to widespread brain areas so that multiple structures are simultaneously informed about any given local effect.[2] Relatedly, in the reverse direction, global brain activity is made available to individual local circuits by electrical oscillations; this is often referred to as "top-down" control.[3] The second fundamental feature of the brain is its persistent activity; that is, the ability of an input to induce and maintain a long-lasting activity trace long after the input has already vanished, even during sleep.[4] Electrical oscillations appear to facilitate these functions via their capacity to coordinate groups of neurons and to divide information into transmittable chunks.

The collective electrical activity of the neurons of the brain is such that signals from large populations of neurons can be

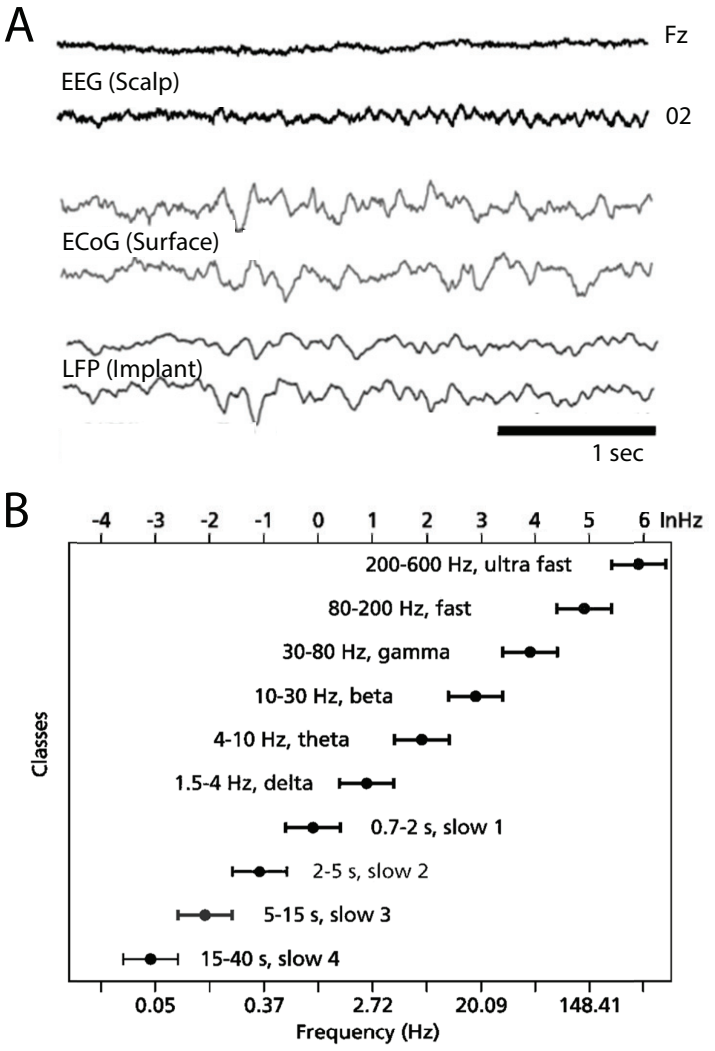
recorded, either with high fidelity from electrodes inside the tissue of the brain (local field potential recording, or LFP), or in an attenuated form from outside the head (through electroencephalography, or EEG). Both during sleep and in waking states, the LFP and EEG show perpetually changing activity (see Figure 1A). Sometimes large-amplitude slow oscillations are predominant, while at other times small-amplitude fast oscillations are present, but most often many rhythms coexist simultaneously.

Neuronal oscillations have been found to exist in the brains of all animals. In mammals, electrical signals over a broad range of frequencies from as low as one wave every forty seconds (0.025 Hz) to as high as six hundred waves per second (600 Hz) have been recorded. In addition, perhaps the best documented but least emphasized fact about brain dynamics is that spectral features of the EEG and LFP are similar in all mammals, independent of brain size. Every known EEG pattern of the human brain is present in all other mammals investigated to date. Furthermore, the correlations of various families of frequencies with aspects of overt behavior and cognition both within and across species have led to the idea of frequency bands: groups of oscillation frequencies in the brain that act as single functional entities (for example, all frequencies from 5 – 8 Hz may act similarly; Figure 1B). Scientists have classified at least ten mutually interacting oscillation bands that are mainly defined by their behavioral correlations. The rhythms constantly interact with each other and form a linear progression on a natural logarithmic scale (Figure 1B again).[5] The fact that these oscillations are highly organized and evolutionarily conserved leads to hypotheses about their function: oscillations may enable neurons to form "assemblies" and synchronize enough to effectively propagate information in neural networks. Second, an even more oscillation-centric interpretation is that synchronization of various brain nuclei is the embodiment of perception, and that since oscillations would be essential to synchronization, they are the key to perception.[6]

Most forms of brain rhythms result from rhythmic inhibitory synaptic transmission (inputs from neurons with net inhibitory effects on the other neurons around them) onto bulk neuronal populations including the information-carrying excitatory neurons.[7] The rhythmic inhibitory volleys from these cell populations provide windows of alternating reduced (inhibited) and enhanced excitability and offer natural temporal frames for grouping or temporally "chunking" neuronal activity into what appear to be functionally related groups of action potentials. The neurons generating these grouped action potentials are called *cell assemblies* and constitute the basic units of information processing. This stop-start parsing function of neuronal oscillators (and their hierarchical cross-frequency coupling organization, detailed below) can support "syntactical" rules for neural communication that are known to both sender and receiver, making communication more straightforward than if areas of the brain had to interpret long uninterrupted messages or stochastic patterns of action potentials.

In addition to instantaneously organizing neurons and inducing them to fire action potentials together, oscillations may play broader and more task-specific organizational roles. Functionally, many different oscillations often co-occur in the same brain state and can interact with each other either within the same brain structure or across anatomical regions. The nature of these interactions of oscillations is hierarchical: the phase of the slower oscillation modulates the power of the faster ones, a mechanism known as cross-frequency

Figure 1
Electrical Oscillations in the Brain



(A) Recordings of brain waves occurring over approximately three seconds. Each line is a recording from one electrode with abcissa representing time and ordinate representing voltage. Top two lines are recorded from outside the skull (EEG), the middle two lines are recordings from inside the skull but on the brain surface (ECoG; electrocorticography), and the bottom two are recorded from electrodes inside the brain. (B) Illustration of the families, or "bands," of oscillatory rhythms in the brain; each is labeled with a horizontal bar. Note that a system of rhythms is formed with a logarithmic relationship among the constituent oscillations. Source: (A) courtesy of Gregory Worrell of the Mayo Clinic and Scott Makeig of the University of California, San Diego; (B) from György Buzsáki and Andreas Draguhn, "Neuronal Oscillations in Cortical Networks," *Science* 304 (2004): 1926–1929.

coupling.[8] An illustration of the effect of brain rhythms on neural communication is the interaction between (5 – 8 Hz) "theta" rhythm in the hippocampus and the higher-frequency (30 – 90 Hz) "gamma" rhythm in the neocortex.[9] With theta oscillations, the hippocampus can temporally coordinate and re-route inputs from other cortical regions (during exploratory behavior, for example) so that the incoming information from disparate regions arrives at approximately the same time and at the phase when the receiver hippocampus is most able to process it.[10]

Oscillations also appear to participate in another major task of the brain: learning new information in order to effectively shape future action. Brains, small and large, are predictive devices that exploit the recurrence of events to learn and use effective actions for various future situations. Learning and memory allow the brain to evolve and adapt to the constantly changing realities brought into our lives by new places, new social acquaintances, new decisions, new positions, and new roles.

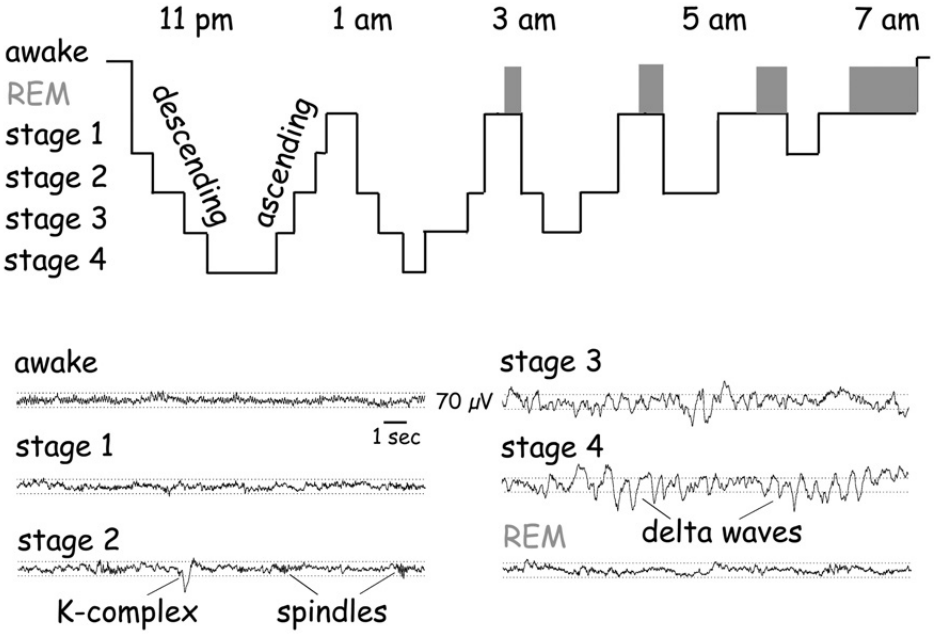
One of the basic tenets of modern neuroscience is that learning and memory are accomplished by the creation or alteration of synaptic connections between neurons. Synchronization of neuronal activity, as occurs during oscillations, can play a key role in the formation of new connections, physically connecting neurons (each carrying information about a different aspect of the world) in order to allow storage of new associations between the elements of the world represented by those neurons. The ability of synapses to strengthen or weaken communication among neurons as a result of experience is called *plasticity*, and it is a cardinal mechanism for adaptation and survival.

In summary, neuronal oscillations are a syntactical structure that is essential to the brain's basic functions of information transmission and computation.[11] Oscillations may also make learning possible by precipitating coordinated changes in internal circuits as life is experienced. With this background, we turn to sleep, focusing on how brain rhythms seem to allow all the tasks of sleep to be accomplished.

Over the last few decades, brain researchers have used empirical evidence to attempt to define the relationships between sleep, learning, and memory. Initial studies showed that although all memories decay with time, they do so more slowly during sleep than during waking. As increasing numbers of studies with larger sample sizes began to show similar results, researchers widely accepted that new experiences may interfere with earlier memories, and that sleep may be a "temporary shelter" in which memories can persist better than during waking.[12]

Later views, however, began to incorporate the knowledge that sleep is not a singular entity but rather is composed of two distinct electrochemical substates known as slow-wave sleep (SWS or non-REM) and rapid eye movement (REM) sleep, and that these physiologically distinct states may play separate roles in memory. During a given episode of sleep, these states appear in a cyclical and relatively stereotyped pattern (Figure 2). Sleep begins with a light form of SWS, progresses to deeper SWS (during which time it is more difficult to awaken the individual and the slow-wave electrical activity is more powerful), retreats back to shallow SWS, and finally concludes with REM sleep before beginning a new cycle. As mentioned, SWS is characterized by large slow waves, which occur at 0.5 to 4 Hz and are quite distinct from waking rhythms (Figure 2, bottom right); in contrast, local field potential recordings of REM sleep look very similar to those of the waking state, with smaller-amplitude gamma waves dominating the

Figure 2
Stages of Sleep



Top panel is a graph of the depth of sleep (depth greater and arousability less toward bottom of the graph) showing a cyclic alternation between SWS and REM sleep. At bottom are example tracings for each state. Note the difference in brain wave amplitude and frequency across states. Source: György Buzsaki, *Rhythms of the Brain* (New York: Oxford University Press, 2006).

neocortex and theta-nested gamma waves (gamma-wave packets occurring at regular portions or phases of theta cycles) in the hippocampus. Chemically, SWS and REM states are also distinct: SWS correlates with a clear decrease in the activity of brain systems secreting the neuromodulators serotonin, histamine, and acetylcholine; but REM sleep involves the selective reinstatement of waking-like acetylcholine system activity. Additionally, SWS and REM each involve the activation of unique regions in the brain stem.

REM sleep initially grabbed the most attention in the scientific community, and a number of researchers found strong correlations between pre-sleep learning and subsequent REM sleep duration.[13] Furthermore, increased REM sleep during human nighttime sleep predicted better performance in later procedural tasks.[14] Animals such as rats also showed greater memory retention and behavioral performance on memory-requiring tasks after sleep with increased REM.[15] Deprivation of REM sleep in animals seemed to impair memory for complex tasks, but not simpler tasks.[16] In humans, the story is not as clear, with only certain complex tasks and procedure-related tasks yielding consistent positive associations with REM.[17] Moreover, REM sleep increases in individuals with major depressive disorder, but

at least in geriatric populations, memory is clearly impaired rather than improved during the depressive episode.[18] It is not clear whether the REM increase causes the memory change in this case, but it is notable that successful pharmacologic treatment of depression decreases or almost entirely eliminates REM.

REM is associated with some of the most recognizable and fascinating aspects of the experience of sleep, however. In experiments where human subjects were awakened during sleep, REM was associated with the most vivid, bizarre dreams, while SWS was associated with more realistic-seeming dreams.[19] Additionally, in a study where people were awakened during REM, they made more associations between ideas and were more able than usual to solve complex anagrams and other problems.[20] On the other hand, SWS has emerged as a critical part of sleep that is capable of changing synaptic weights and that has been tied to a greater degree to "declarative" memories: memories of consciously declarable facts and events.

A simple yet persuasive model of sleep suggests that during the day the brain processes information and coordinates perception and action, while nighttime serves for proper maintenance of the entire system. In 1982, psychopharmacologist Alexander Borbély formally proposed that sleep has a homeostatic function: the regulation of a component he called "S," which builds with waking (causing "sleep pressure") and is relieved or dissipated during sleep.[21] Support for this view can be found in experiments showing that the magnitude of slow waves is larger at the start of sleep and weaker later in sleep; it also becomes larger yet with sleep deprivation, as if it responded to the degree of need for sleep.[22] Additional support comes from data showing that when a particular region of brain is used intensely prior to sleep, slow-wave amplitude is selectively larger in that region.[23] Sleep may play an even more general function by effectively removing potentially neurotoxic waste products that accumulate in the central nervous system during waking states.[24]

Recently, the hypothetical S-process was linked with synaptic connections via an influential theory stating that synaptic connectivity is the factor that builds as waking experience prolongs.[25] Under this theory, synaptic connections are constantly built as new associations are made during the course of waking. As more and more associations are made, the brain may build too many, running the risk of becoming less functional. Sleep might scale back synapses to allow preserved brain functioning. The homeostatic model specifically states that sleep weakens each individual synapse by a universal proportion across all synapses, although it does not attribute specific roles for REM and non-REM stages. Using advanced microscopic and molecular biological techniques to directly view the anatomical synaptic connections in mice, neuroscientist Giulio Tononi and his colleagues in fact did demonstrate that sustained waking results in more synapses and sustained sleep brings with it decreasing numbers of synapses.[26] However, they also demonstrated that some synapses were actually created during sleep – a result that requires explanation.

Somewhat at odds with the theory that sleep performs a universal synapse-reducing role is the view that sleep specifically participates in memory consolidation. The conflict comes from the notion that to consolidate memories, synapses active during waking learning experience must be selectively kept active or even enhanced, which is in apparent opposition to the synapse-weakening homeostatic hypothesis.

Over the past two decades, numerous experiments have been performed that

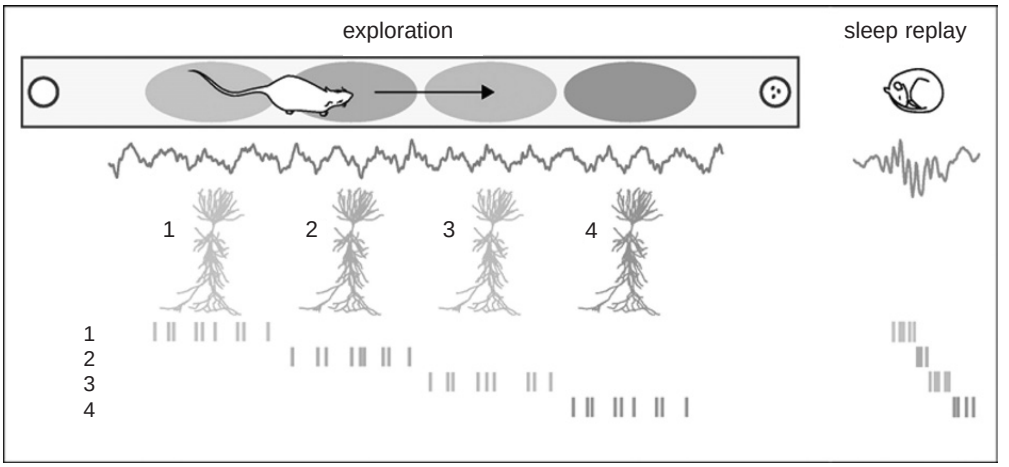
support a "two-stage model" of memory consolidation: during sleep – that is, after waking acquisition – memories are not wiped away or simply made to decay less slowly, but are often actually improved, molded, and shaped.[27] Procedural memories, such as learned finger-tapping rhythms, can be sped up or even abstracted from one hand to another during sleep.[28] Imagination-based mental practice of movements or observation of others can also lead to improvements after sleep.[29] All of this supports the idea that new skills can be built, gained, or incorporated into the proper brain sites by sleep, which appears to be performing an active and constructive role. Additionally, people may rearrange their knowledge after a session of sleep, as evidenced by their committing novel errors in a systematic fashion that demonstrates they have internalized a general concept rather than the precise details of a given set of pre-sleep events.[30] Finally, there appears to be some emotional or even conscious control over which memories will be improved during sleep: telling human subjects that they will receive rewards for retaining certain information seems to be enough to cause memory for that information to be the most bolstered by sleep.[31]

The seemingly opposing views of synaptic homeostasis and consolidation may be better approached if the synaptic perspective is supplemented with an electrical rhythm–based approach to categorizing and studying sleep. Oscillatory rhythms can be used to divide sleep into SWS and REM. Furthermore, in animal models, advanced neuroscientific tools can aid in exploring neural activity in novel ways, such as recording action potential activity from many neurons simultaneously during behavior and sleep.

The key player in the memory consolidation process is the sharp-wave ripple (SPW-R): a brief (50 – 150 ms) electrical rhythm generated by an intrinsic self-organizing process in the hippocampus, which apparently provides a perfect mechanism for the precise consolidation of waking experience.[32] This brief rhythm is cross-frequency-coupled with other rhythms such as slow waves and sleep spindles, and it represents the most synchronous physiological pattern in the mammalian brain: 10 to 18 percent of all neurons in the hippocampus and highly interconnected regions (subiculum and entorhinal cortex) discharge during these events. SPW-Rs occur during "offline" brain states, such as waking immobility and SWS; apparently, they also represent a counterpart to the theta oscillation, which is present during movement, active waking learning, and REM. SPW-Rs have been hypothesized as an ideal mechanism for transferring information and inducing synaptic changes,[33] especially since artificially induced SPW-R–like patterns can strengthen synapses even in brain slices *in vitro*.[34] Another piece of evidence supporting SPW-Rs' key role in information transfer is the fact that sequences of neuronal assemblies present during waking behavior are replayed (and at higher speed) during subsequent SPW-R events (see Figure 3).[35]

The replay phenomenon during SPW-Rs may be a direct mechanism for consolidation, given mounting evidence showing that the more often two neurons fire together (as in SPW-Rs), the stronger the synapse between them becomes.[36] This theory states that when neurons fire repeatedly with a particular timing relative to each other, the synapses between them will strengthen, knitting them together into a cohesive representation of a given percept. This may be precisely what happens when the brain replays waking activity patterns through SPW-Rs occurring in slow-wave sleep. Direct evidence for

Figure 3
Schematic Replay of Waking Neuronal Activity during Sleep



At left: during waking states, experience in the environment leads to certain sequences of neuronal firing, as in this example where a rat has a number of neurons fire in a sequence corresponding to places the animal has visited. At right: fast replay of the same firing sequence of neuronal activation during a sharp wave ripple in sleep. Experiments have shown that replay after waking experience is greater than prior to waking experience. Source: adapted with permission from Gabrielle Girardeau and Michaël Zugaro, "Hippocampal Ripples and Memory Consolidation," *Current Opinion in Neurobiology* 21 (3) (2011): 452 – 459.

SPW-Rs' causal role in memory consolidation comes from studies in which SPW-Rs were selectively eliminated after learning.[37] Such targeted interference did not affect other aspects of sleep but produced a large impairment in learning.

While SPW-Rs can emerge solely from a self-organized process, they are often biased to occur by other brain-wide oscillations, including sleep states. SPW-Rs are modulated by thalamo-cortical sleep-spindle oscillations (12 – 16 Hz);[38] both SPW-Rs and spindles are modulated by slow oscillations;[39] and finally, each of these three rhythms is modulated by the ultraslow (0.1 Hz) oscillation.[40] Each of these rhythms predominates in certain regions of the brain and their co-modulation may allow for a coordination of the regions they reside in to accomplish a complex task: two-stage memory consolidation. Specifically, we currently believe that memories initially stored in the hippocampus during waking, and later, during sleep consolidation, are transferred to other regions such as the cerebral cortex for more permanent storage (though they leave a trace in the hippocampal system).[41] So it may be that during SWS, slow waves, which are known to originate in the cortex, entrain the hippocampus and SPW-Rs to transmit information when the cortex is ready to receive it.[42] This would be a direct (though anatomically opposite) sleeping analog to the waking theta rhythm of the hippocampus entraining the cortex so that the cortex sends information when the hippocampus is ready to receive it. Recent evidence supports this scenario with data indicating that synaptic connections

in the cortex can be formed particularly well during the receptive phase of the slow-wave oscillation.[43]

In a set of creative experiments, neuroscientist Jan Born and colleagues were able to demonstrate the importance of slow-wave activity by experimentally enhancing slow oscillations in sleeping people using transcranial electrical stimulation outside the scalp.[44] They showed that increased power of slow oscillations leads to increased memory gain upon waking. It remains to be demonstrated whether sleep spindles and slow oscillations contribute to memory consolidation by their own mechanisms or by their entrainment of hippocampal SPW-Rs.[45]

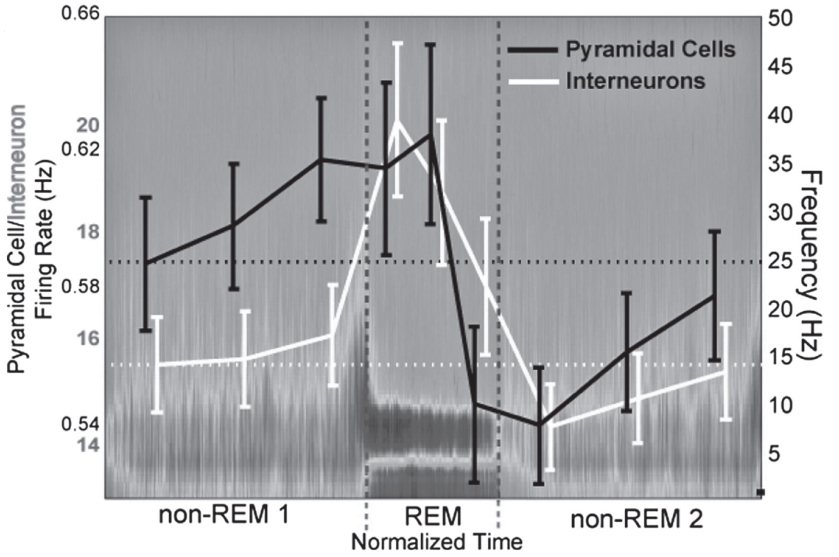
To summarize, selective and time-compressed reactivation of learning-induced firing patterns is present during sleep. Selective interference with the key rhythms underlying the replay of spike sequences impairs memory performance, whereas enhancement of the relevant oscillatory patterns improves memory performance. It should be pointed out, however, that while the experiments discussed above demonstrate the vital importance of SPW-Rs and other rhythms, they do not provide direct information about the mechanism by which those rhythms bring about change, since artificial perturbations affect the dynamic of neuronal interactions at many levels.

An obvious limitation of both the homeostasis and consolidation models is that they are mute on the role of REM sleep and do not account for the complex choreography of the cyclic SWS-REM sleep process. One place for insight is the temporal dynamics of these processes during sleep: while the homeostasis model assumes that synaptic weights generally undergo a monotonic decrease over the entire duration of sleep, consolidation experiments demonstrate that compressed sequence replay of waking firing patterns vanishes after thirty to sixty minutes.[46] Remarkably, this time corresponds to the onset of the first REM episode in rodents. Thus, it is possible that REM and SWS play different but complementary roles in the sleep-memory process. Recent research suggests that this may indeed be the case.

While prior evidence showed that overall firing rates of both excitatory and inhibitory neurons decreased steadily during sleep (as predicted by the homeostasis model), new findings show that most of the rate decrease could be accounted for by a number of brief REM episodes (see Figure 4).[47] It is assumed that firing rate may correlate with overall neuronal synaptic connectivity. These experiments show that firing rates actually increase slightly during SWS, which occupies the majority of sleep, but that during brief REM bouts there are dramatic drops in firing rate that cumulate over multiple REM episodes. Thus, while several hours of waking is necessary to reach the hypothesized saturation of synaptic weights and firing rates, short REM episodes with electrophysiological characteristics of the waking brain can paradoxically bring about rapid downscaling of global firing rates. These results bring us back to the critical importance of both SWS and REM and their interaction for allowing sleep to carry out its full purpose of both consolidation and homeostatic synaptic downscaling.

The data presented above instruct us that an approach oriented toward measuring and understanding the electrical rhythms of the brain gives a fuller picture of the function and mechanisms of sleep. We have proposed a possible fusion of the two dominant models of sleep through the division of labor by REM and SWS. In its simplest form, SWS may be in charge of consolidation and selective enhancement

Figure 4
Complementary Roles for SWS and REM in Neuronal Physiology



In two types of neurons studied (pyramidal cells and inhibitory interneurons), SWS led to a slight increase of spiking activity and possibly synaptic connectivity (rising slopes), while during REM sleep, action potential generation and possibly synaptic connectivity decreased (falling slopes). Source: A. D. Grosmark, K. Mizuseki, E. Pastalkova, K. Diba, and G. Buzsaki, "REM Sleep Reorganizes Hippocampal Excitability," *Neuron* 75 (2012): 1001 – 1007.

of learning-related neuronal firing patterns, whereas REM may be responsible for homeostatic downscaling of rates and synaptic weights. Sleep always begins with SWS; thus, consolidation and strengthening of important synapses come first, followed by a hypothetically "evenhanded" downscaling of synapses by REM. This would ensure that the most important synapses are not lost in a population of relatively similarly-sized synapses. The process then can swing back to SWS to consolidate the content modified by REM before preparing for another iteration of the cycle.

Alternatively, one can point out the overlapping aspects of the models. The consolidation model is de facto homeostatic because selective enhancement of learning-related connectivity can occur only at the expense of other synapses. This is due to the fact that brain-wide synaptic weights (and, by extension, brain-wide firing rates) must remain constant; otherwise, epileptic activity would result.[48] Therefore, any increase must be accompanied by a compensatory decrease.

New findings may also need to be considered when evaluating the theories discussed here. First, only a small fraction of neurons active during SPW-R events are the same as those that were active during prior learning, and many more sequences occur than could be accounted for by the sequences observed during waking behavior.[49] Several experiments show that the spike content of SPW-Rs does not exclusively relate to the activity during the most recently experienced situation or to the most frequent neuronal sequences of the

preceding waking period.[50] Finally, new evidence is accumulating that suggests SPW-Rs have a constructive (and not just post-hoc) role in learning: it has been observed that groups of neurons that fire in a specific sequence during a novel running behavior will have actually fired in that same order (though compressed in time) in sleep *before* the new behavior occurs.[51] This suggests that the brain may use pre-constructed network structures during behavior, and that sleep plays a role in the pre-behavior activity of those network constructs.

It is, of course, likely that the model presented above is overly simplistic and misses many details. Additionally, certain synapses and neurons may have varying roles in network function that may accordingly cause them to be treated differently by sleep, and many assumptions about the nature of these systems may need to be re-addressed in light of new data. However, it is worth seriously considering the theory that sleep is a general tool used by mammals to tune their entire neural system to be able to properly acquire, select, and store information. If experience is gained during waking, then the connections formed by the registration of that experience are not only passively sheltered by sleep, but are amplified, reorganized, and generalized (by consolidation); additionally, the slate is wiped mostly clean and many synapses reset for the next day of learning (by synaptic downscaling).

Without sleep, the waking brain system might be encumbered by having to take on additional roles, such as preventing the formation of too many connections and selecting only the most useful to remain. It seems nature has found that allowing the waking state to be fully dedicated to tasks such as perception, learning, and motor control is most adaptive as long as sleep will come a few hours later to clean and selectively reorganize the system. Thus, the system of adaptation, learning, and memory may owe its efficiency to sleep's balancing effects.

Additional evidence for sleep's crucial role in proper adaptive functioning comes from cases of sleep deprivation. Acute sleep deprivation can lead to seizures, impaired cognition, poor memory, mood lability, irritability, and even frank psychosis with disorganized thinking and poor ability to accurately perceive reality.[52] Indeed, there is a well-documented but poorly understood link between many neuropsychiatric disorders (including major depressive disorder, anxiety disorders, and bipolar disorder) and sleep, and there is an emerging consensus that sleep disorganization might be causally related to the cognitive-affective problems associated with these disorders.

In summary, sleep is a pair of special modes in the brain: SWS and REM. Both of these states are closed to outside inputs and work in tandem to prepare and clean the learning and memory system while retaining important information for later use – all so that the brain can be maximally focused on learning and adapting to its surroundings during the next period of awake activity.

ENDNOTES

* Contributor Biographies: BRENDON O. WATSON is a clinical psychiatrist and a research fellow at Weill Cornell Medical College at Cornell University and is doing postdoctoral research work at the Buzsáki Lab at the New York University School of Medicine. His research interests include sleep mechanisms and emotional processing in animal models and his clinical interests include affective and personality disorders. He has published in such journals as *Dialogues in Clinical Neuroscience*, *Frontiers in Neuroscience*, *Frontiers in Neural Circuits*, and *Neuron*.

GYÖRGY BUZSÁKI is the Biggs Professor of Neural Sciences at the New York University School of Medicine. His laboratory's research goal is to investigate syntactical structures that enable internal communication within the brain. He is among the top 1 percent of the most cited authors in neuroscience, the recipient of the 2011 Brain Prize, and the author of *Rhythms of the Brain* (2006). He also sits on the editorial board of numerous journals, including *Science* and *Neuron*.

1 Björn Rasch and Jan Born, "About Sleep's Role in Memory," *Physiological Reviews* 93 (2013): 681 – 766.

2 Stanislas Dehaene and Jean-Pierre Changeux, "Experimental and Theoretical Approaches to Conscious Processing," *Neuron* 70 (2011): 200 – 227; and G. Tononi and G. M. Edelman, "Consciousness and Complexity," *Science* 282 (1998): 1846 – 1851.

3 Francisco Varela, Jean-Philippe Lachaux, Eugenio Rodriguez, and Jacques Martinerie, "The Brainweb: Phase Synchronization and Large-Scale Integration," *Nature Reviews Neuroscience* 2 (2001): 229 – 239; and Andreas K. Engel, Pascal Fries, and Wolf Singer, "Dynamic Predictions: Oscillations and Synchrony in Top-Down Processing," *Nature Reviews Neuroscience* 2 (2001): 704 – 716.

4 György Buzsáki, *Rhythms of the Brain* (New York: Oxford University Press, 2006).

5 György Buzsáki and Andreas Draguhn, "Neuronal Oscillations in Cortical Networks," *Science* 304 (2004): 1926 – 1929.

6 C. M. Gray and W. Singer, "Stimulus-Specific Neuronal Oscillations in Orientation Columns of Cat Visual Cortex," *Proceedings of the National Academy of Sciences* 86 (1989): 1698 – 1702.

7 G. Buzsáki, L. W. Leung, and C. H. Vanderwolf, "Cellular Bases of Hippocampal EEG in the Behaving Rat," *Brain Research* 287 (1983): 139 – 171; György Buzsáki and James J. Chrobak, "Temporal Structure in Spatially Organized Neuronal Ensembles: A Role for Interneuronal Networks," *Current Opinion in Neurobiology* 5 (1995): 504 – 510; M. A. Whittington, R. D. Traub, N. Kopell, B. Ermentrout, and E. H. Buhl, "Inhibition-Based Rhythms: Experimental and Mathematical Observations on Network Dynamics," *International Journal of Psychophysiology* 38 (2000): 315 – 336; and Nikos K. Logothetis, "The Underpinnings of the BOLD Functional Magnetic Resonance Imaging Signal," *The Journal of Neuroscience* 23 (2003): 3963 – 3971.

8 Anatol Bragin et al., "Gamma (40 – 100 Hz) Oscillation in the Hippocampus of the Behaving Rat," *The Journal of Neuroscience* 15 (1995): 47 – 60; Charles E. Schroeder and Peter Lakatos, "Low-Frequency Neuronal Oscillations as Instruments of Sensory Selection," *Trends in Neurosciences* 32 (2009): 9 – 18; and G. Buzsáki and B. O. Watson, "Brain Rhythms and Neural Syntax: Implications for Efficient Coding of Cognitive Content and Neuropsychiatric Disease," *Dialogues in Clinical Neuroscience* 14 (2012): 345 – 367.

9 Anton Sirota et al., "Entrainment of Neocortical Neurons and Gamma Oscillations by the Hippocampal Theta Rhythm," *Neuron* 60 (2008): 683 – 697.

10 Ibid.; and György Buzsáki, "Neural Syntax: Cell Assemblies, Synapsembles, and Readers," *Neuron* 68 (2010): 362 – 385.

11 Buzsáki, "Neural Syntax: Cell Assemblies, Synapsembles, and Readers."

12 Rasch and Born, "About Sleep's Role in Memory"; and Jeffrey M. Ellenbogen, Jessica D. Payne, and Robert Stickgold, "The Role of Sleep in Declarative Memory Consolidation: Passive, Permissive, Active or None?" *Current Opinion in Neurobiology* 16 (2006): 716 – 722.

[13] William Fishbein, Chris Kastaniotis, and Dennis Chattman, "Paradoxical Sleep: Prolonged Augmentation Following Learning," *Brain Research* 79 (1974): 61 – 75.

[14] Stefan Fischer, Manfred Hallschmid, Anna Lisa Elsner, and Jan Born, "Sleep Forms Memory for Finger Skills," *Proceedings of the National Academy of Sciences* 99 (2002): 11987 – 11991.

[15] W. Wetzel, D. Balschun, S. Janke, D. Vogel, and T. Wagner, "Effects of CLIP (Corticotropin-Like Intermediate Lobe Peptide) and CLIP Fragments on Paradoxical Sleep in Rats," *Peptides* 15 (1994): 237 – 241.

[16] Chester A. Pearlman, *Effect of Rapid Eye Movement (Dreaming) Sleep Deprivation on Retention of Avoidance Learning in Rats* (Fort Belvoir, Va.: Defense Technical Information Center, 1969), http://archive.rubicon-foundation.org/xmlui/handle/123456789/8608.

[17] Rasch and Born, "About Sleep's Role in Memory."

[18] Charles F. Reynolds III et al., "EEG Sleep in Elderly Depressed, Demented, and Healthy Subjects," *Biological Psychiatry* 20 (1985): 431 – 442.

[19] J. A. Hobson and Edward F. Pace-Schott, "The Cognitive Neuroscience of Sleep: Neuronal Systems, Consciousness, and Learning," *Nature Reviews Neuroscience* 3 (2002): 679 – 693.

[20] Denise J. Cai, Sarnoff A. Mednick, Elizabeth M. Harrison, Jennifer C. Kanady, and Sara C. Mednick, "REM, not Incubation, Improves Creativity by Priming Associative Networks," *Proceedings of the National Academy of Sciences* 106 (2009): 10130 – 10134.

[21] Alexander A. Borbély, "A Two-Process Model of Sleep Regulation," *Human Neurobiology* 1 (1982): 195 – 204.

[22] Vladyslav V. Vyazovskiy et al., "Cortical Firing and Sleep Homeostasis," *Neuron* 63 (2009): 865 – 878.

[23] Tadanobu Yasuda, Kyo Yasuda, Richard A. Brown, and James M. Krueger, "State-Dependent Effects of Light-Dark Cycle on Somatosensory and Visual Cortex EEG in Rats," *American Journal of Physiology: Regulatory, Integrative, and Comparative Physiology* 289 (2005): R1083 – R1089.

[24] Lulu Xie et al., "Sleep Drives Metabolite Clearance from the Adult Brain," *Science* 342 (2013): 373 – 377.

[25] Giulio Tononi and Chiara Cirelli, "Sleep Function and Synaptic Homeostasis," *Sleep Medicine Reviews* 10 (2006): 49 – 62; and Vladyslav V. Vyazovskiy and Kenneth D. Harris, "Sleep and the Single Neuron: The Role of Global Slow Oscillations in Individual Cell Rest," *Nature Reviews Neuroscience* 14 (2013): 443 – 451.

[26] Stephanie Maret, Ugo Faraguna, Aaron B. Nelson, Chiara Cirelli, and Giulio Tononi, "Sleep and Waking Modulate Spine Turnover in the Adolescent Mouse Cortex," *Nature Neuroscience* 14 (2011): 1418 – 1420.

[27] Hiroyuki Miyamoto and Takao K. Hensch, "Reciprocal Interaction of Sleep and Synaptic Plasticity," *Molecular Interventions* 3 (2003): 404 – 417; and Joel H. Benington and Marcos D. Frank, "Cellular and Molecular Connections between Sleep and Synaptic Plasticity," *Progress in Neurobiology* 69 (2003): 71 – 101.

[28] Gais et al., "Early Sleep Triggers Memory for Early Visual Discrimination Skills."

[29] Ysbrand D. Van Der Werf, Els Van Der Helm, Menno M. Schoonheim, Arne Ridderikhoff, and Eus J. W. Van Someren, "Learning by Observation Requires an Early Sleep Window," *Proceedings of the National Academy of Sciences* 106 (2009): 18926 – 18930; and Ursula Debarnot, Thomas Creveaux, Christian Collet, Julien Doyon, and Aymeric Guillot, "Sleep Contribution to Motor Memory Consolidation: A Motor Imagery Study," *Sleep* 32 (2009): 1559 – 1565.

[30] Henry L. Roediger and Kathleen B. McDermott, "Creating False Memories: Remembering Words not Presented in Lists," *Journal of Experimental Psychology: Learning, Memory, and Cognition* 21 (1995): 803 – 814.

31 S. Fischer and J. Born, "Anticipated Reward Enhances Offline Learning During Sleep," *Journal of Experimental Psychology : Learning, Memory, and Cognition* 35 (2009) : 1586 – 1593.

32 György Buzsáki, "Two-Stage Model of Memory Trace Formation : A Role for 'Noisy' Brain States," *Neuroscience* 31 (1989) : 551 – 570.

33 Ibid. ; and J. J. Chrobak and G. Buzsáki, "Selective Activation of Deep Layer (V – VI) Retrohippocampal Cortical Neurons During Hippocampal Sharp Waves in the Behaving Rat," *The Journal of Neuroscience* 14 (1994) : 6160 – 6170.

34 György Buzsáki, Helmut L. Haas, and Edmund G. Anderson, "Long-Term Potentiation Induced by Physiologically Relevant Stimulus Patterns," *Brain Research* 435 (1987) : 331 – 333.

35 M. A. Wilson and B. L. McNaughton, "Reactivation of Hippocampal Ensemble Memories During Sleep," *Science* 265 (1994) : 676 – 679 ; Mattias P. Karlsson and Loren M. Frank, "Awake Replay of Remote Experiences in the Hippocampus," *Nature Neuroscience* 12 (2009) : 913 – 918 ; and György Buzsáki and Fernando Lopes da Silva, "High Frequency Oscillations in the Intact Brain," *Progress in Neurobiology* 98 (3) (2012) : 241 – 249, http ://www.ncbi.nlm.nih .gov/pubmed/22449727.

36 Donald O. Hebb, *The Organization of Behavior* (New York : Wiley, 1949).

37 Karlsson and Frank, "Awake Replay of Remote Experiences in the Hippocampus" ; and Gabrielle Girardeau, Karim Benchenane, Sidney I. Wiener, György Buzsáki, and Michaël B. Zugaro, "Selective Suppression of Hippocampal Ripples Impairs Spatial Memory," *Nature Neuroscience* 12 (2009) : 1222 – 1223.

38 Athanassios G. Siapas and Matthew A. Wilson, "Coordinated Interactions Between Hippocampal Ripples and Cortical Spindles During Slow-Wave Sleep," *Neuron* 21 (1998) : 1123 – 1128 ; and Anton Sirota, Jozsef Csicsvari, Derek Buhl, and György Buzsáki, "Communication Between Neocortex and Hippocampus During Sleep in Rodents," *Proceedings of the National Academy of Sciences* 100 (2003) : 2065 – 2069.

39 M. Steriade, D. Contreras, R. Curró Dossi, and A. Nuñez, "The Slow (< 1 Hz) Oscillation in Reticular Thalamic and Thalamocortical Neurons : Scenario of Sleep Rhythm Generation in Interacting Thalamic and Neocortical Networks," *The Journal of Neuroscience* 13 (1993) : 3284 – 3299.

40 Anton Sirota et al., "Communication Between Neocortex and Hippocampus During Sleep in Rodents" ; Matthias Mölle, Oxana Yeshenko, Lisa Marshall, Susan J. Sara, and Jan Born, "Hippocampal Sharp Wave-Ripples Linked to Slow Oscillations in Rat Slow-Wave Sleep," *Journal of Neurophysiology* 96 (2006) : 62 – 70 ; and David Sullivan et al., "Relationships Between Hippocampal Sharp Waves, Ripples, and Fast Gamma Oscillation : Influence of Dentate and Entorhinal Cortical Activity," *The Journal of Neuroscience* 31 (2011) : 8605 – 8616.

41 Buzsáki, "Two-Stage Model of Memory Trace Formation : A Role for 'Noisy' Brain States."

42 Anton Sirota and György Buzsáki, "Interaction Between Neocortical and Hippocampal Networks via Slow Oscillations," *Thalamus & Related Systems* 3 (2005) : 245 – 259 ; and Thomas G. Hahn, Bert Sakmann, and Mayank R. Mehta, "Differential Responses of Hippocampal Subfields to Cortical Up-Down States," *Proceedings of the National Academy of Sciences* 104 (2007) : 5169 – 5174.

43 Peter B. Kruskal, Lucy Li, and Jason N. MacLean, "Circuit Reactivation Dynamically Regulates Synaptic Plasticity in Neocortex," *Nature Communications* 4 (2013), doi :10.1038/ncomms3574.

44 Lisa Marshall, Halla Helgadóttir, Matthias Mölle, and Jan Born, "Boosting Slow Oscillations During Sleep Potentiates Memory," *Nature* 444 (2006) : 610 – 613.

45 Alain Destexhe and Terrence Sejnowski, *Thalamocortical Assemblies : How Ion Channels, Single Neurons, and Large-Scale Networks Organize Sleep Oscillations* (Oxford : Oxford University Press, 2001), 452.

46 Wilson and McNaughton, "Reactivation of Hippocampal Ensemble Memories during Sleep."

47 A. D. Grosmark, K. Mizuseki, E. Pastalkova, K. Diba, and G. Buzsáki, "REM Sleep Reorganizes Hippocampal Excitability," *Neuron* 75 (2012): 1001 – 1007.

48 Ronald J. Racine, "Modification of Seizure Activity by Electrical Stimulation: II. Motor Seizure," *Electroencephalography and Clinical Neurophysiology* 32 (1972): 281 – 294.

49 David J. Foster and Matthew A. Wilson, "Reverse Replay of Behavioural Sequences in Hippocampal Place Cells During the Awake State," *Nature* 440 (2006): 680 – 683; Kamran Diba and György Buzsáki, "Forward and Reverse Hippocampal Place-Cell Sequences During Ripples," *Nature Neuroscience* 10 (2007): 1241 – 1242; and Zoltán Nádasdy, Hajime Hirase, András Czurkó, Jozsef Csicsvari, and György Buzsáki, "Replay and Time Compression of Recurring Spike Sequences in the Hippocampus," *The Journal of Neuroscience* 19 (1999): 9497 – 9507.

50 David Dupret, Joseph O'Neill, Barty Pleydell-Bouverie, and Jozsef Csicsvari, "The Reorganization and Reactivation of Hippocampal Maps Predict Spatial Memory Performance," *Nature Neuroscience* 13 (2010): 995 – 1002; Margaret F. Carr, Shantanu P. Jadhav, and Loren M. Frank, "Hippocampal Replay in the Awake State: A Potential Substrate for Memory Consolidation and Retrieval," *Nature Neuroscience* 14 (2011): 147 – 153; and Anoopum S. Gupta, Matthijs A. A. van der Meer, David S. Touretzky, and A. David Redish, "Hippocampal Replay is Not a Simple Function of Experience," *Neuron* 65 (2010): 695 – 705.

51 Gupta et al., "Hippocampal Replay is Not a Simple Function of Experience"; and G. Dragoi and S. Tonegawa, "Preplay of Future Place Cell Sequences by Hippocampal Cellular Assemblies," *Nature* 469 (2011): 397 – 401.

52 Louis Jolyon West, Herbert H. Janszen, Boyd K. Lester and Floyd S. Cornelisoon, Jr., "The Psychosis of Sleep Deprivation," *Annals of the New York Academy of Sciences* 96 (1962): 66 – 70.

A Hard Scientific Quest: Understanding Voluntary Movements

Emilio Bizzi & Robert Ajemian

Abstract: In this article we explore the complexities of what goes on in the brain when one wishes to perform even the simplest everyday movements. In doing so, we describe experiments indicating that the spinal cord interneurons are organized in functional modules and that each module activates a distinct set of muscles. Through these modules the central nervous system has found a simple solution to controlling the large number of muscle fibers active even during the execution of the simplest action. We also explore the many different neural signals that contribute to pattern formations, including afferent information from the limbs and information of motor memories.

EMILIO BIZZI, a Fellow of the American Academy since 1980 and President of the Academy from 2006 to 2009, is Institute Professor in the Department of Brain and Cognitive Sciences and Investigator at the McGovern Institute for Brain Research at the Massachusetts Institute of Technology.

ROBERT AJEMIAN is a Research Scientist at the McGovern Institute for Brain Research at the Massachusetts Institute of Technology.

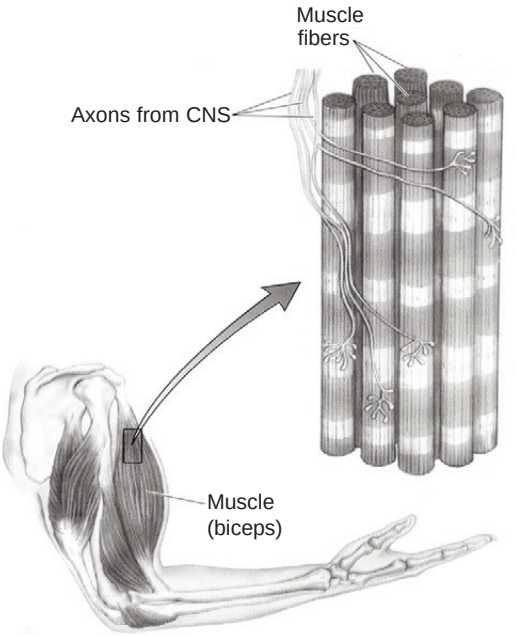
*(*See endnotes for complete contributor biographies.)*

Scientists and nonscientists alike rarely stop to consider what is going on in their brains when they perform a voluntary movement such as reaching for an object, throwing a ball, or driving a car. Why? Presumably they may realize that translating something as evanescent as a wish to move into muscle contractions must be an awfully complicated process. Indeed, they are right: the neural processes that subserve even the simplest everyday actions are incredibly complex and only partially understood. In this essay we take up the challenge of explaining what we know about this fascinating and complex topic.

Let us begin with the basic fact that, in general, our movements – even the simplest actions – are accomplished through activation of a large number of muscles. For example, if you are sitting at your desk typing at your computer and decide to turn to pick up a cup of coffee, you will activate, approximately at the same time, the eye muscles, the numerous muscles in the neck, and the muscles of the shoulder, arm, forearm, and fingers. A simple computation would show that your brain has activated at least thirty muscles. But note that each muscle is made up of cells called muscle fibers, and that each muscle fiber receives a neural input via its own nerve fiber (see Figure 1). It follows that the number of elements

doi:10.1162/DAED_a_00324

Figure 1
Muscle Fibers and Axons from Motor Neurons



Axons are the long threadlike part of a nerve cell along which impulses are conducted from the cell body to other cells. Source: *Frontiere Della Vita* (Rome: Istituto Della Enciclopedia Italiana, 1999).

controlled by the neural motor system is very large, even during simple movements.

Imagine now what must be taking place in the brain of an athlete in the heat of a soccer match, when practically all the muscles of the body must be precisely coordinated, with little time for preplanning. Clearly, the soccer player trying to score a goal has neither the time nor the inclination to explicitly formulate the command signals to control the millions and millions of muscle fibers in his or her body, and must instead rely on an effective simplifying strategy. How our brains cope with this inherent complexity remains one of the fundamental questions in motor system neuroscience.

Years ago a group of neuroscientists, including one of the authors of this essay, decided to investigate this basic question by launching a series of exploratory searches aimed at identifying the way in which the central nervous system (CNS; the complex of nerve tissues that control the activities of the body, comprising the brain and the spinal cord) controls the multitude of muscle fibers that are activated during movements.[1] We started by focusing on the spinal cord in lower vertebrates and quickly found that a special group of cells called interneurons – neurons that transmit impulses between other neurons, and that are interposed between the sensory portion of the spinal cord and its motor output – are the key elements that implement the simplifying strategy.[2]

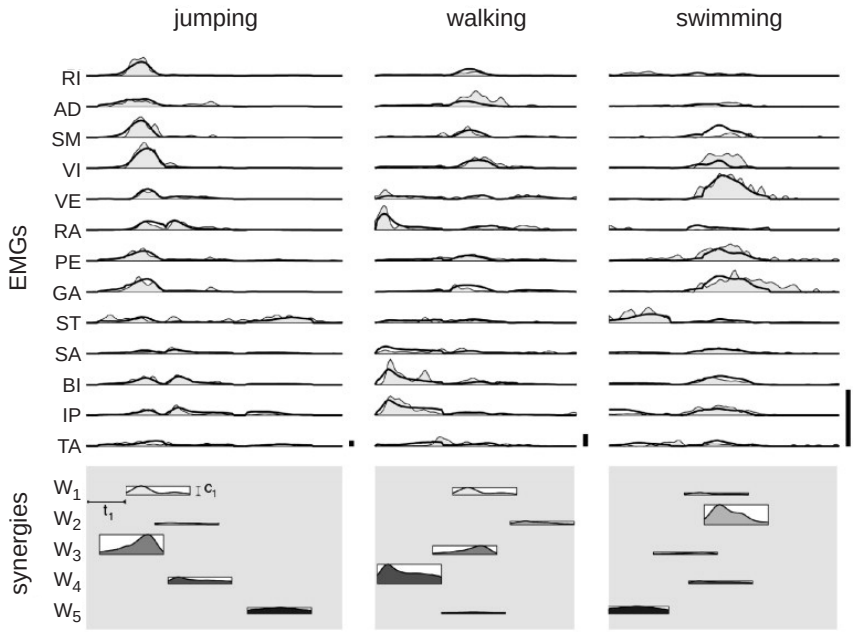
Interneurons are organized in functional modules, and each module activates a particular set of muscles in distinct proportions. We labeled this entity of patterned muscle activation a *muscle synergy*. This modular spinal structure is the central piece of a discrete combinatorial system that utilizes a finite number of discrete elements (that is, the muscle synergies) to express a voluntary movement. The combinatorial system is controlled by the neurons residing in the cortical frontal areas (the cerebral cortex covering the frontal lobe). Anatomically, the cortical neurons transmit impulses to select and combine spinal modules. Following the arrival of cortical command signals, a cascade of neural events ensues: the activated spinal modules fire the motoneurons (nerve cells forming part of a pathway along which impulses pass from spinal cord to a muscle) and their motor nerves induce a depolarization of the muscle fibers, which in turn is followed by muscle contraction and movements. Researchers can easily record the electromyographic activity (EMG) – the depolarization of muscle fibers – with electrodes placed in or on the muscle surface.

To identify the muscle synergies we used a factorization algorithm that takes as input all of the muscle EMG data and extracts from these data both a set of generative muscle synergies and a coefficient of each synergy during the composition of a particular motor behavior. The experimental evidence supporting the idea that the CNS uses muscle synergies as output modules is illustrated in Figure 2. This illustration shows that a small number of synergies explain a large fraction in the variation of muscle patterns. In other words, not as many individual muscle functions need to be controlled as one might have initially thought.

Figure 2 shows the EMG records for frogs that are jumping, walking, and swimming. The upper (unshaded) section of Figure 2 lists the names and the EMGs of thirteen leg muscles. The shaded wave areas in this section of the figure represent the rectified, filtered, and integrated EMGs recorded during the execution of a single instance of jumping, walking, and swimming movement. The thick line defining the contour of the wave represents the outcome of a computation that reconstructs the muscle patterns by utilizing the muscle synergies extracted by the factorization procedure. The lower (shaded) section of Figure 2 shows the coefficients of the five synergies that were found through the factorization. The coefficients are placed in a rectangular box whose width corresponds with synergy duration; their position indicates onset delay and the height represents amplitude of EMG. Figure 2 demonstrates two important points: 1) the same synergies are found to contribute to different movements (note that synergies W_1, W_3, and W_4 are a constituent of both jumping and walking, but with different coefficients of activation of EMGs; the synergy W_5 is used in both jumping and swimming); and 2) different behaviors may be constructed by linearly combining the same synergies with different timing and scaling factors. Recent results from the study of muscle patterns during a variety of movements in humans, monkeys, and other vertebrates have shown that combining a small set of muscle synergies appears to be a general strategy that the CNS utilizes for simplifying the control of limb movements.

The specific factorization algorithm that we used to extract the underlying synergies from the overall EMG data set is known as the nonnegative matrix factorization.[3] Other factorization algorithms could have been used, such as the popular principal component analysis (PCA); but neuroscientist Matt Tresch and his colleagues have shown that whatever technique one uses to identify the synergies, the end results are

Figure 2
Synergies and Variations of Muscle Patterns



The main muscles of synergy W_1 are RI, AD, PE, and GA. The main muscles of synergy W_2 are SM, VE, PE, and GA. The main muscles of synergy W_3 are RI, SM, and VI. The main muscles of synergy W_4 are RA, BI, and IP. The main muscles of Synergy W_5 are ST and IP. These synergies were extracted by pooling together the EMGs of three frogs during jumping, walking, and swimming movements. Source: Emilio Bizzi, Vincent C. K. Cheung, Andrea d'Avella, Philippe Saltiel, and Matthew Tresch, "Combining Modules for Movement," *Brain Research Reviews* 57 (2008): 125–133.

essentially the same.[4] This suggests that the observed muscle synergies are real, as opposed to an artifact of the data analysis. Additional observations corroborate the independent physiological existence of muscle synergies as fundamental and irreducible units of motor control that are linearly combined to generate movement.[5]

The experimental evidence described above indicates that the peripheral sections of the motor system operate as a discrete combinatorial system. In a way, then, the motor system is like language, a system in which discrete elements and a set of rules for combining them can generate a large number of meaningful entities that are distinct from those of their elements. Thus, we may have solved the problem of how the motor system copes with having to control so many different muscles and motor units during the course of a movement: it does so through intelligent modularization at the level of the spinal cord.

But having proposed a solution for one problem, we are immediately led to another: how does the brain figure out the correct combinations of synergies that are required to execute a motor act? Certainly, what is impressive about the motor system is its capacity to find original motor solutions to an infinite set of ever-changing circumstances. This capacity is entirely de-

pendent upon the computations performed by neural circuitries of the cortical areas of the frontal lobe (the lobes on each cerebral hemisphere lying immediately behind the forehead). These cortical areas generate signals that combine, select, and activate the spinal modules. Understanding these computations has been the main goal of neuroscientists, neurologists, and psychologists involved in the study of motor control. Some progress toward this goal has been made, but as we will discuss, many of the hard questions remain unanswered.

Our description of the way in which cortical commands generate patterns of activity for activating the spinal cord modules will begin by considering the major inputs and outputs of the motor cortical regions.

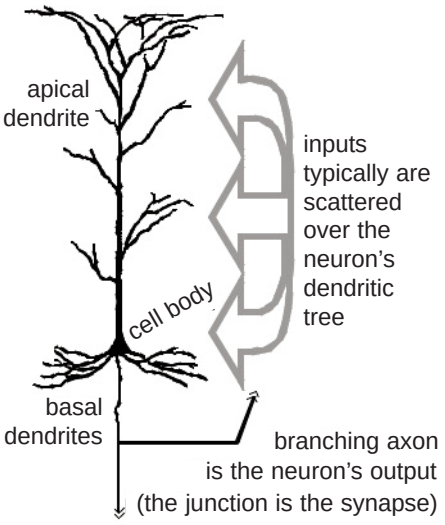
In each frontal lobe hemisphere, there are at least four major regions concerned with generating signals for voluntary movements: the dorsal and ventral premotors (the cortical areas in front of the motor cortex), the supplementary motor area, and the primary motor cortex.[6] These highly interconnected regions receive diverse modalities of information (inputs) from a variety of sources, including: 1) external sensory information about the state of the world (such as visual, auditory, tactile information); 2) internal sensory information about the state of the body (such as muscle length and tendon force); 3) the executive attentional system for determining behavioral saliency; and 4) inputs from major subcortical areas such as the cerebellum and the basal ganglia (whose roles in motor control are somewhat obscure). These signals are conveyed to the motor cortical areas where they connect to the dendritic tufts of the large output cells of the cortical layers 5 and 6 (see Figure 3).[7] There these signals are somehow integrated into a coherent unit to set up a neuronal depolarization, which is conveyed via the dendritic tree to the cell body of the big output cells in layer 5, and from there via long pathways to the spinal cord.

There are a variety of output pathways made of axons of cortical layers 5 and 6 that connect premotor and primary motor cortical areas with a different class of spinal neurons. One of these descending pathways conveys information about an impending movement, an observation suggesting that it may be part of a cortico-spinal circuit contributing to an early shaping of motor commands. Additional output pathways include: 1) cortico-spinal fibers terminating at the level of the interneurons, which activate the spinal cord modules and therefore are responsible for the expression of the muscle synergies;[8] 2) a cortico-motoneuronal pathway from the most caudal sections of the primary motor cortex,[9] though we do not yet know how these two descending sets of fibers cooperate in the execution of voluntary movements; 3) an important set of fibers connecting the motor cortex to the basal ganglia; and 4) a set of fibers connecting the cortex to the cerebellum in the recurrent loop, which create a cerebellar pathway whose complex function is possibly related to reentry circuits that contribute to shaping the construction of cortical patterns of activity. (The function of the cerebellum has long been a source of debate possibly relating to the function of cortical patterns activity; see Figure 4.)

But also critical is that the descending pathways are mirrored by ascending sets of fibers forming numerous reentry circuits. Thus, the intimate ties between the cortex and the periphery are an essential feature of the "system" for movement.

There is a vast amount of data indicating that the motor cortex plays a central role in generating motor behavior, but there is lack of consensus on how neural process-

Figure 3
Pyramidal Neuron

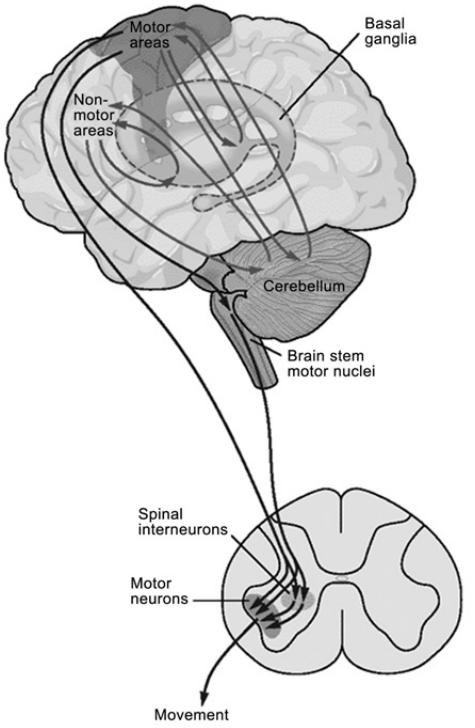


Source: *Frontiere Della Vita* (Rome: Istituto Della Enciclopedia Italiana, 1999).

ing within the cortical areas of the frontal lobe contribute to voluntary movements. An approach to interpret cortex neural activity, first introduced by Edward Evarts at the National Institutes of Health, was based on recording the activity of single cortical neurons in monkeys and then correlating their firing rate with joint motion, force, and limb posture.[10] Evarts concluded that the motor cortical neurons likely encoded the muscular force that is needed for movement. In the early 1980s, neuroscientist Apostolos Georgopoulos, using a modified behavioral setup, showed that cortical neurons recorded from the primary motor cortex were broadly tuned to the direction of hand movements.[11] This correlation suggests that the motor cortex encodes high-level parameters of movement such as direction in task space, rather than low-level parameters such as muscle forces. But the story did not end there. In the last few decades, researchers have implicated the motor cortex in the encoding of a litany of different movement variables, including hand velocity, hand position, joint angles, joint torques, movement sequence information, and movement curvature. Further, the response properties appear nonstationary, changing with behavioral contexts and choice of task.[12] So what does this all mean? What does the motor cortex really do?

To put all these observations in perspective, we should consider the limitations of microelectrode recordings. In acute recording, one or a handful of neurons are recorded. In chronic recording, an array of electrodes is implanted that can record from roughly one hundred neurons simultaneously. Yet there are millions of neurons in the motor cortices, each one highly interconnected to other motor cortical neurons and to the many input sources that project to it. The endeavor is thus lim-

Figure 4
The Cortex, Cerebellum, Brain Stem, and a Section through the Spinal Cord



The motor cortical areas are shown with some of the fibers connecting it to the basal ganglia and the cerebellum. Two cortical spinal pathways are shown, one a direct pathway from the cortex to the spinal motorneurons and another connecting the motor cortex to the spinal interneurons. Source: Eric R. Kandel, James H. Schwartz, and Thomas M. Jessell, eds., *Principles of Neural Science*, 4th ed. (New York: McGraw-Hill, 2000).

ited both by problems of undersampling (you are only recording from a tiny fraction of a population) and problems of limited sampling information (usually you do not know in what layer a recorded neuron resides, meaning you do not know where in the context of this complicated and highly distributed circuit your neuron fits). An analogy: suppose that you know nothing about how computers work, but are asked to figure it out by sticking a volt meter into different regions of a computer while it runs various programs, recording its electric potential. Considering this challenge, perhaps it is not so surprising that the manner in which patterns form at the output layer of the motor cortex to generate movements remains a mystery.

Clearly, to move our understanding of the motor cortex forward, new approaches are needed. In the last few years, prompted by the urge to look for new avenues, neurophysiologists and computational neuroscientists have joined forces in order to make sense of the neuronal recording data and then generate theories and models of the motor cortex. The most notable proposed models, such as optimal feedback control and recurrent neural network, are attempts at formulating a comprehensive motor control theory;[13] but because their focus is based on our limited knowl-

edge of neurons, the resulting models had only a modest impact on the field. In general, these models failed to consider the complex interactions among different classes of cortical cells and the role of the recurrent circuits that link the cortex with the spinal cord, basal ganglia, and cerebellum. The cortical neurons' activity should have been evaluated in a different way, since these cells belong to an ensemble.

Fortunately, thanks to developments in molecular biology, a panoply of new techniques is becoming available. For example, imaging techniques are being developed that will enable an experimenter to record from thousands of neurons simultaneously, while at the same time monitoring the anatomical changes taking place within the circuit at the synaptic level. Different strains of viruses carrying channelrhodopsin into targeted populations of neurons now make it possible to activate/inhibit neural circuitry. This is an important development because neural circuits are inherently parallel and highly interconnected, meaning that it is difficult to understand what one part of the circuit is doing in isolation of the rest of the circuit. Such technological breakthroughs, together with the development of mathematical tools for processing and modeling high-dimensional distributed dynamical systems, may change the playing field in systems neuroscience in the years to come.

Neuronal activity in the motor cortical areas is a complex function of sensory inputs, regional and local interactions among cells, and cortical reentry circuits. In addition, recent investigations have established that one more significant way to fire the cortical neurons is just to image an action without producing an actual movement. There is also evidence that shifting from one mental task to another changes the pattern of brain activation. These results were obtained by monitoring regional blood flow either with a positron emission tomograph or fMRI (functional magnetic resonance imaging). It is of interest that most of the activation was found in the premotor and supplementary areas, but less so for the primary cortex for both motor imagery as well as movement observation. These findings show that the deliberate representations of actions involve activation similar to those occurring during voluntary movements. These observations are important as they expand the scope of the motor system beyond the generation of actions. Imaging the brain during a motor action means to evaluate the consequence of self-programmed movements before execution, while also providing ways to represent other people's actions.

As a consequence of the intense exploration of the motor imaging underlying physiology is the realization that cells in the premotor and primary motor cortex are active when a subject plans an action. This finding has opened the way to record from the human cortex and utilize the neural signals for prosthetic devices.

Closely linked to the motor imaging of actions is the brain's representation of motor memories; that is, when we learn a skill, how is that skill represented in our neural circuits? Since most of what we do is guided by what we have learned, this capacity for motor learning embodies a crucial facet of our existence. Imagine how difficult life would become if every time we engaged in a routine act, like tying our shoes, we had to perform it with the skill level of a novice. Instead, as we go through life we gain facility and acquire expertise in the form of motor memories: memories of how to perform skilled motor acts. Where are these motor memories stored and how are they represented?

In the case of computers, we know where and how information is stored. It is stored in the digital switches of transistors that

are housed and addressed in specific locations within the computer (the hard drive or random-access memory) or various external digital media (such as a disk). In the case of human declarative memory – which, loosely, is factual information about one's life (names, places, events, and so on) – we also have a fairly good idea. Declarative memory is stored in the medial temporal lobe of the brain (a region of the cerebral cortex that is located beneath the lateral fissure on both cerebral hemispheres of the mammalian brain) including the hippocampus (the elongated ridges on the floor of each lateral ventricle of the brain, thought to be the center of emotion, memory, and the autonomic nervous system), the entorhinal cortex (an area of the brain located in the medial temporal lobe that functions as a hub in a widespread network for memory and navigation), and perirhinal cortex (a region of the medial temporal lobe of the brain that receives highly processed sensory information from all sensory regions, and is generally accepted to be an important region for memory). Motor memory, on the other hand, appears to be broadly distributed across all of the major components of the motor circuit, including the motor cortices, the cerebellum, and the basal ganglia; and how motor memories are stored still remains murky, though we have clues.

One clue has come from force field studies in which monkeys adapted their reaching movements to different environments and perturbations while the activity of single neurons in their motor cortices was recorded.[14] As expected, when a monkey learned to move in a novel context, the activity patterns of the recorded neurons changed. This finding is generally consistent with the synaptic trace theory of memory, which says that memories are embodied in patterns of synaptic connections (synapses are the junctions between two nerve cells consisting of a minute gap across which impulses pass by diffusion of a neurotransmitter) that change in an experience-dependent fashion, such that after the experience, the circuit is capable of generating a new output. However, it was unexpectedly found that some of the neurons maintained their altered activity patterns even when the animal stopped performing the new task and returned to the original task. Further, as the behavior switched from one task to another, the pattern of activity of the neurons changed in an unpredictable fashion.

In a similar vein, a pair of studies used the technique of two-photon microscopy to study anatomical changes in the synaptic connectivity of the mouse motor cortex during the learning of new motor skills.[15] This remarkable technology allows experimenters to visualize individual synaptic spines, the smallest unit of information transmission in the brain, over periods of weeks or months. As expected, these studies showed that learning is indeed accompanied by the formation of new synaptic spines. However, these studies showed something else that was unexpected, even shocking: when the animals are not learning anything, the synaptic spines are still turning over at a high rate. In fact, the rate of turnover is so high in the baseline conditions that most of the new spines created during the formation of the new memory will be gone in a matter of months or, at most, a couple of years. Yet these same motor memories are known to persist for the animal's entire life.

These observations lead to a profound paradox. If we believe that memories are made of patterns of synaptic connections sculpted by experience, and if we know, behaviorally, that motor memories last a lifetime, then how can we explain the fact that individual synaptic spines are constantly turning over and that aggregate synaptic strengths are constantly fluctuating? How can the memories outlast their puta-

tive constitutive components? This is currently one of the great mysteries in motor neuroscience and, in fact, all of systems neuroscience, reinforced by the dozens of two-photon microscopy studies that have found that regardless of which region of the cortex is examined, the synapses are constantly turning over. How is the permanence of memory constructed from the evanescence of synaptic spines?

In an attempt to answer this question, we recently developed a new type of neural network with the distinguishing feature of synapses that are constantly changing even when no learning is taking place.[16] We showed that under certain conditions – conditions that hold during motor learning – the network can stably learn a variety of skills despite these constant weight changes. The basic point of the model is to reexamine the notion of what constitutes a memory. Neural circuits are highly redundant in that there are many more synapses than there are neurons, with each neuron being contacted, on average, by ten thousand synaptic spines. Thus, within a neural circuit, many different configurations of synapses can give rise to the same input-output processing. In other words, a network can perform the same function even if its synapses undergo change. With this in mind, we suggest that a memory, instead of being composed by a fixed pattern of synaptic weights, is actually embodied by a fixed pattern of input/output processing at the level of neural activities. This flexibility gives the system some slack to accommodate synaptic turnover, an inevitable fact of cell biology, since synapses are made of proteins, which have short lifetimes. Models, like this one, that rely on the stochastic dynamics of complex systems may prove to be fertile territory for understanding recent and future data on synaptic dynamics.

A final question is whether models like ours can simultaneously shed light on both the small-scale physiology/anatomy and the larger-scale behavior of a subject. After all, the ultimate goal of neuroscience is to mechanistically link the physical entity of the brain to the more ethereal phenomena of the mind. This is not always easy to do and often a model is constructed to explain data in one domain or the other, but not both. Here we have found that a model with perpetually fluctuating synaptic connections may explain an interesting result from the kinesiology and sports science community that has been known for over one hundred years.

The finding is called the warm-up decrement and it can be illustrated as follows: to function at a peak performance level, a professional athlete trained in a fine motor skill must practice or warm up for an extended period of time immediately prior to performing. For example, professional golfers and professional tennis players will practice for an hour or more before playing in a major competition. In a certain sense, this seems strange. These athletes have spent much of their lives practicing a particular skill, so why do they still need to practice for so long before competing? A robot that performs a skill needs only to be turned on and shortly thereafter it will execute the skill to the best of its abilities. So why do human experts need additional practice right before performance?

One possibility is that the practice is needed to warm up the athlete's muscles, ligaments, and tendons. However, this theory has been refuted by experiments in which the body is warmed up by other means, resulting in sub-peak performance. Further, athletes of all calibers widely accept that the warm-up ought to use the same skill set that will be used in performance. If a professional tennis player were to practice squash an hour before playing a tennis match, the results would be disastrous, though many of the same muscles are used in both activities. So what, then,

is going on during this period of warm-up? One explanation could be a need for continuous neural recalibration to optimize performance if, as proposed in our model, synapses are always changing. Thus, over time, there might be a slight drop-off in performance on the basis of synaptic turnover alone. For an expert who performs at the highest level, even a slight decrement in performance can be obvious and exploited by competition, thereby making practice immediately before an event required to fine-tune the network into a state of optimal performance.

Whether or not the proposed model is correct in its explanation remains to be determined. But what we know for sure is that the continued use of modern imaging technology to probe synaptic dynamics will provide crucial data in the years ahead to constrain and inform our efforts at understanding the neural processes that underlie motor learning.

In this review we have focused on the hard scientific questions involved in understanding the seemingly effortless generation of voluntary movements. With respect to the peripheral motor system (spinal cord and muscles), we have pointed out the many difficulties associated with controlling millions of muscle fibers partitioned across dozens of muscles, and described how, through spinal cord modularity, the CNS has found a simplifying solution. However, no answers have yet been found to explain how the cortical motor areas of the frontal lobe construct the spatiotemporal patterns of neural activity necessary to activate the spinal cord, enabling it to execute a specific movement. Certainly, we do know that high-level movement goals and attention-related signals are represented in the premotor areas and that the spread of these signals to the primary motor cortex, possibly already primed by afferent information about limb posture, will somehow trigger the retrieval of motor memories and, subsequently, the formation of a signal to the spinal cord. But the detail of this complicated process, which critically involves coordinate and variable transformations from spatial movement goals to muscle activations, needs to be elaborated further. Phrased more fancifully, we have some idea as to the intricate design of the puppet and the puppet strings, but we lack insight into the mind of the puppeteer.

We have also discussed the hard problem of where and how motor memories are stored. This, too, is a difficult and unsolved problem, in large measure because of the highly distributed and interconnected nature of neural circuits. Based on first principles, we can be sure that memory storage in the brain, however it works, will differ radically from information storage in a computer. New computational paradigms may be needed to provide a greater understanding, and we have here briefly described one model that attempts to shift the paradigm based on our knowledge of perpetually fluctuating synapses.

On the face of it, investigating the problems of both pattern formation (for the purpose of control) in the motor cortex and motor memory storage in the aggregate motor circuit is going to be a daunting affair requiring the combined efforts of physiologists, molecular biologists, and computational neuroscientists. But as the history of science has shown, it is possible that nature might have developed some surprisingly simple and unexpected shortcuts that, if discovered, may go a long way toward providing the ultimate answers. After all, who would have guessed that a simple receptive field in the visual cortex might be the key to recognizing the complexities of a face?

ENDNOTES

* Contributor Biographies: EMILIO BIZZI, a Fellow of the American Academy since 1980 and President of the Academy from 2006 to 2009, is Institute Professor in the Department of Brain and Cognitive Sciences and Investigator at the McGovern Institute for Brain Research at the Massachusetts Institute of Technology. His recent publications include articles in such journals as *Frontiers in Computational Neuroscience*, *Proceedings of the National Academy of Sciences*, *Neuron*, and the *Journal of Neurophysiology*.

ROBERT AJEMIAN is a Research Scientist at the McGovern Institute for Brain Research at the Massachusetts Institute of Technology. His publications include articles in such journals as *Neuron*, *Cerebral Cortex*, the *Journal of Motor Behavior*, and the *Journal of Neurophysiology*.

Authors' Note: We would like to acknowledge the following grants, which funded research that contributed to this essay: National Science Foundation Grant IIS-0904594, Program for Collaborative Research in Computational Neuroscience; and National Institutes of Health Grants (NINDS) NS44393 and NS068103.

1 Emilio Bizzi and Vincent C. K. Cheung, "The Neural Origin of Muscle Synergies," *Frontiers in Computational Neuroscience* 7 (51) (2013): 1, doi:10.3389/fncom.2013.00051.

2 Emilio Bizzi, Vincent C. K. Cheung, Andrea d'Avella, Philippe Saltiel, and Matthew Tresch, "Combining Modules for Movement," *Brain Research Reviews* 57 (2008): 125 – 133.

3 Ibid.

4 Matthew C. Tresch, Vincent C. K. Cheung, and Andrea d'Avella, "Matrix Factorization Algorithms for the Identification of Muscle Synergies: Evaluation on Simulated and Experimental Data Sets," *Journal of Neurophysiology* 95 (4) (2006): 2199 – 2212, doi:10.1152/jn.00222.2005.

5 Bizzi and Cheung, "The Neural Origin of Muscle Synergies."

6 The cortex, the outer layer of the cerebrum, is composed of folded gray matter and plays an important role in consciousness.

7 A dendrite is a short branched extension of a nerve cell, along which impulses received from other cells at synapses are transmitted to the cell body.

8 Ariel J. Levine, Christopher A. Hinckley, Kathryn L. Hilde, Shawn P. Driscoll, Tiffany H. Poon, Jessica M. Montgomery, and Samuel L. Pfaff, "Identification of a Cellular Node for Motor Control Pathways," *Nature Neuroscience* 17 (4) (2014): 586 – 593, doi:10.1038/nn.3675.

9 Jean-Alban Rathelot and Peter L. Strick, "Subdivisions of Primary Motor Cortex Based on Cortico-Motoneuronal Cells," *Proceedings of the National Academy of Sciences* 106 (3) (2009): 918 – 923, doi:10.1073/pnas.0808362106.

10 E. V. Evarts, "Relation of Pyramidal Tract Activity to Force Exerted During Voluntary Movement," *Journal of Neurophysiology* 31 (1) (1968): 14 – 27.

11 Apostolos P. Georgopoulos, John F. Kalaska, Roberto Caminiti, and Joe T. Massey, "On the Relations between the Direction of Two-Dimensional Arm Movements and Cell Discharge in Primate Motor Cortex," *The Journal of Neuroscience* 2 (11) (1982): 1527 – 1537.

12 Stephen H. Scott, "Inconvenient Truths about Neural Processing in Primary Motor Cortex," *The Journal of Physiology* 586 (5) (2008): 1217 – 1224, doi:10.1113/jphysiol.2007.146068

13 For the optimal feedback control model, see Emanuel Todorov and Michael I. Jordan, "Optimal Feedback Control as a Theory of Motor Coordination," *Nature Neuroscience* 5 (11) (2002): 1226 – 1235, doi:10.1038/nn963; and for the recurrent neural network model, see Mark M. Churchland, John P. Cunningham, Matthew T. Kaufman, Justin D. Foster, Paul Nuyujukian, Stephen I. Ryu, and Krishna V. Shenoy, "Neural Population Dynamics during Reaching," *Nature* 487 (7405) (2012): 51 – 56.

14 Chiang-Shan Ray Li, Camillo Padoa Schioppa, and Emilio Bizzi, "Neuronal Correlates of Motor Performance and Motor Learning in the Primary Motor Cortex of Monkeys Adapting to an External Force Field," *Neuron* 30 (2) (2001): 593 – 607.

15 Tonghui Xu, Xinzhu Yu, Andrew J. Perlik, Willie F. Tobin, Jonathan A. Zweig, Kelly Tennant, Theresa Jones, and Yi Zuo, "Rapid Formation and Selective Stabilization of Synapses for Enduring Motor Memories," *Nature* 462 (2009): 915 – 919, doi:10.1038/nature08389; and Guang Yang, Feng Pan, and Wen-Biao Gan, "Stably Maintained Dendritic Spines are Associated with Lifelong Memories," *Nature* 462 (2009): 920 – 924, doi:10.1038/nature08577.

16 Robert Ajemian, Alessandro D'Ausilio, Helene Moorman, and Emilio Bizzi, "A Theory for How Sensorimotor Skills are Learned and Retained in Noisy and Nonstationary Neural Circuits," *Proceedings of the National Academy of Sciences* 110 (52) (2013): E5078 – E5087, doi:10.1073/pnas.1320116110; and Robert Ajemian, Alessandro D'Ausilio, Helene Moorman, and Emilio Bizzi, "Why Professional Athletes Need a Prolonged Period of Warm-Up and Other Peculiarities of Human Motor Learning," *Journal of Motor Behavior* 42 (6) (2010): 381 – 388.

Feelings: What Are They & How Does the Brain Make Them?

Joseph E. LeDoux

Abstract: Traditionally, we define "emotions" as feelings and "feelings" as conscious experiences. Conscious experiences are not readily studied in animals. However, animal research is essential to understanding the brain mechanisms underlying psychological function. So how can we make study mechanisms related to emotion in animals? I argue that our approach to this topic has been flawed and propose a way out of the dilemma: to separate processes that control so-called emotional behavior from the processes that give rise to conscious feelings (these are often assumed to be products of the same brain system). I will use research on fear to explain the way that I and many others have studied fear in the laboratory, and then turn to the deep roots of what is typically called fear behavior (but is more appropriately called defensive behavior). I will illustrate how the processes that control defensive behavior do not necessarily result in conscious feelings in people. I conclude that brain mechanisms that detect and respond to threats non-consciously contribute to, but are not the same as, mechanisms that give rise to conscious feelings of fear. This distinction has important implications for fear and anxiety disorders, since symptoms based on non-conscious and conscious processes may be vulnerable to different factors and subject to different forms of treatment.

JOSEPH E. LEDOUX, a Fellow of the American Academy since 2006, is the Henry and Lucy Moses Professor of Science at New York University. He is the author of the books *Synaptic Self: How Our Brains Become Who We Are* (2002), *The Emotional Brain* (1996), and *Anxious: Using the Brain to Better Understand and Treat Fear and Anxiety* (forthcoming July 2015). He has published research in such publications as *Nature*, *Neuron*, and *The Journal of Neuroscience*, and was elected to the National Academy of Sciences in 2013.

The human mind has two fundamental psychological motifs. Descartes's proclamation, "I think, therefore I am,"[1] illustrates one, while Melville's statement, "Ahab never thinks, he just feels, feels, feels," exemplifies the other.[2] Our Rationalist inclinations make us want certainty (objective truth), while the Romantic in us basks in emotional subjectivity. Psychology and neuroscience recognize this distinction: cognition and emotion are the two major categories of mind that researchers study. But things were not always quite like this.

Rational thought has always been treated as a product of the mind, and emotions were traditionally viewed as belonging to the body.[3] Descartes, following Plato's lead, said that humans differ from animals in that we have a rational mind (a soul), but are similar to animals in that we have bodily passions (emotions) that interfere with reason. Somewhere along the way, two things happened to give us our more

doi: 10.1162/DAED_a_00319

integrated modern view. First, we began to consider human emotions as mental states; and second, many began to attribute mental states, including both thoughts and feelings, to animals. Darwin, for example, viewed emotions as "states of mind," some of which are shared by both humans and other animals.[4] Today, emotions are commonly conceived of as mental states that are felt (consciously experienced) when well-being is affected in some way: we feel fear when threatened, anger when frustrated, joy when things go well, sadness when we lose a friend or loved one, and compassion when we see someone suffer.

If we assume that emotions are feelings, that feelings are states of consciousness, and that states of consciousness are inner private experiences predicated on the awareness of one's own mental activity, questions arise about the scientific study of the brain mechanisms underlying emotions in animals. Emotions and other states of consciousness can – within limits and with some difficulty – be studied in humans, but the study of consciousness in animals is, to put it mildly, challenging.[5] At the same time, due to ethical and technical limits of investigations in humans, work on detailed brain mechanisms of emotion depends on animal research. How do we get around this stumbling block?[6]

Behaviorism, which flourished in the first half of the twentieth century, is a school of thought in psychology that rejects the study of conscious experience in favor of objectively measurable events (such as responses to stimuli). Due to behaviorism's influence, researchers interested in emotion in animals have tended to take one of two approaches. Some have treated emotion as a brain state that connects external stimuli with responses.[7] These researchers, for the most part, viewed such brain states as operating without the necessity of conscious awareness (and therefore as separate from feelings), thus avoiding questions about consciousness in animals.[8] Others argued, in the tradition of Darwin, that humans inherited emotional states of mind from animals, and that behavioral responses give evidence that these states of mind exist in animal brains.[9] The first approach has practical advantages, since it focuses research on objective responses of the body and brain, but suffers from the fact that it ignores what most people would say is the essence of an emotion: the conscious feeling. The second approach puts feelings front and center, but is based on assumptions about mental states in animals that cannot easily be verified scientifically.

When I was getting started in my studies of emotion in animals in the mid-1980s, I adopted a third approach to try to get around these problems.[10] I treated *emotions* in terms of essentially non-conscious brain states that connect significant stimuli with response mechanisms, and *feelings* as conscious experiences arising from these non-conscious brain states.[11] My theory, therefore, emphasized the importance of feelings, but I argued that the brain mechanisms that control emotional responses and those that generate conscious feelings are separate. By separating processes that non-consciously detect and respond to significant stimuli from those that create feelings, emotional mechanisms could be studied in animals without having to solve the problem of whether animals feel emotion, while at the same time honoring the importance of feelings in the human mind and brain.

I have used this strategy for many years in my research on fear in animals, focusing on the mechanisms that detect threatening stimuli and orchestrate defensive responses to deal with the danger.[12] The purpose of this strategy has not been to deny that feelings or other states of consciousness exist in animals, but instead to focus re-

search on questions that can be addressed scientifically, regardless of how the animal consciousness debate turns out. In the meantime, since feelings are an essential factor in human mental life and in psychiatric disorders that afflict humans, conscious feelings can and should be studied in humans. Further, because circuits that operate non-consciously to control emotional responses nevertheless contribute to feelings in humans, research on these same circuits in animals is relevant to human feelings.

As useful as the strategy has been, there has always been something awkward about the scientific separation of emotion (a non-conscious response process) from feeling – the conscious experience of emotion. It is messy, since the terms *emotion* and *feeling* are typically used interchangeably in everyday speech. There is no requirement that scientific language conform to lay meaning (and in fact some argue that a language of science will replace lay terms);[13] but when the terms in question are about mental states that we all experience and talk about, it is harder to escape the compelling pull of the vernacular.[14] This is one challenge psychology faces that most other sciences do not.

In an effort to grapple with this terminological uneasiness, I began rethinking the way we use words like emotions and feelings.[15] This led me to consider the natural history of what we have been calling emotion in animals, a journey that led me to conclude that the roots of so-called emotional behavior are so deep, so old, that it makes no sense to use a term like emotion to describe these behaviors in any organism, including humans. The term emotion is so intricately entwined with conscious feelings that to use it in any other way simply invites confusion. Instead of differentiating between emotion and feeling, I stick with the everyday meaning of the terms, using them interchangeably to refer to the mental states that people experience when they face situations in which survival is challenged or enhanced.

In the remainder of this essay I use the emotion of fear as an example of the points I want to make. I first explain the way that I and many others have studied fear in the laboratory, and then turn to the deep roots of what I long called fear behavior (but now call defensive behavior, as explained below). I offer a different way of talking about fear and other mental states.[16] I conclude by discussing my view about what conscious feelings are, how they relate to non-conscious processes, and how they come about in the brain.

A staple in laboratory studies of fear and its underlying brain mechanisms has been a procedure called *Pavlovian fear conditioning*.[17] (I prefer to use the more neutral term *Pavlovian threat conditioning* to circumvent aforementioned problems associated with discussing "fear" in animals.)[18] In threat conditioning, an insignificant stimulus, such as a tone, occurs in conjunction with an aversive stimulus, typically footshock. Through this pairing, the tone by itself eventually acquires the ability to elicit freezing behavior (a defensive anti-predator behavior in which the animal remains motionless for the purpose of avoiding detection or minimizing attack)[19] and supporting physiological responses (such as changes in heart rate, blood pressure, and other autonomic nervous system adjustments) that help the organism cope with the impending danger. Much has been learned about the neural circuits, cells, synapses, and molecules that make it possible for animals to learn in this way, especially through studies of rodents (particularly rats).[20]

One of the advantages of Pavlovian threat conditioning is that it can be used across a wide range of species. Human studies can-

not provide as much detail about neural mechanisms, but have confirmed to a first approximation that the same brain areas and connections discovered in rats also exist in humans.[21] For example, through studies of rats, we know much about the various subregions of the amygdala that receive the tone and shock, integrate them, store a memory of the association, and use that association to control defensive responses.[22] Subregions of the amygdala are not within the imaging resolution of currently available technology for studying the human brain, but we assume that circuits are likely to be similar because the basic circuitry found in rats also exists in nonhuman primates and performs similar functions in rodents and primates.[23]

Invertebrates are also responsive to Pavlovian threat conditioning.[24] While these animals have different neural circuits than vertebrates, they offer advantages for exploring molecular mechanisms related to intracellular signaling cascades and gene expression. Many of the discoveries made in invertebrates have subsequently been confirmed in rodents.[25] And if they apply to rodents, they likely apply to other mammals, including nonhuman primates and humans.

While studies of detailed brain mechanisms are not possible in humans, studies of our species have the distinct advantage of being able to explore conscious states.[26] Still, we must be careful not to confuse feelings with responses elicited by a threat. When threatened by a stimulus created through threat conditioning or by an innate threat, humans have behavioral and autonomic nervous system responses that anticipate the threat and help prepare the body to cope with the danger that may ensue; further, the amygdala is activated.[27] The person may feel fear, but this does not mean that the same brain circuits create feelings of fear.

For example, the amygdala is activated and physiological responses are expressed even to subliminal (non-conscious) presentations of threat stimuli.[28] In these cases, the subjects are not aware of the stimulus and do not report any particular feeling.[29] Amygdala activation thus does not tell us that fear is felt in a human, and certainly does not alert us to fearful feelings in animals. Confusion results because fearful feelings are often correlated with these amygdala-dependent responses. But correlation does not mean causation; we cannot generalize from stimulus-response mechanisms, which occur widely in animal life, to conscious feelings of fear.[30]

That said, amygdala-based and other defensive circuits do contribute indirectly to feelings of fear, but feelings of fear require more than just amygdala-driven responses in the brain and body. My proposal is that all organisms have the ability to detect and respond to threats, but only organisms that can be conscious of their own brain's activities can feel fear.

Laboratory studies of so-called emotional behavior in animals involve tasks that pose challenges to, or opportunities for enhancing, well-being. Stimuli (such as shocks, food, drink, warmth, and sexual stimulation) are used to motivate responses that help the animal either cope with or benefit from the stimuli (prevent or reduce the impact of a shock or give access to food, drink, warmth, or sex). When humans experience these events, we can have feelings of fear (when threatened) or pleasure (when eating, drinking, having sex, or becoming warm after being in the cold). These behaviors and feelings are so intertwined in us that we think of them as one and the same: we often describe the feelings as emotions and the behaviors as emotional behaviors.

When descending the evolutionary tree in search of the origins of these so-

called emotional behaviors, one quickly finds oneself jumping to lower and lower branches, ending up at the base of the trunk and eventually even digging down into the roots of the tree. Every living organism, from the oldest to the most recent, has to do these things to stay alive and pass its genes on to its offspring. Organisms must detect danger, identify and consume nutrients and energy sources, balance fluids by taking in and expelling liquids, thermoregulate, and reproduce. You do these things, but so do the bacterial cells living in your lower intestine.[31]

This realization turns the scientific language of emotion on its head. What are commonly called emotion functions in humans and animals are not emotional functions at all. They do not exist to make feelings. They are survival functions essential for the continued life of the individual or the species.[32] And in humans, survival functions are sometimes – perhaps often – associated with feelings. But the systems that underlie these functions operate independently from feelings in humans. For example, as noted above, the circuits that control so-called fear responses are not themselves the wellspring of feelings of fear. This raises the question of how feelings of fear or other emotions come about.

The problem of understanding feelings is thus reducible to the problem of understanding consciousness. Consciousness is unobservable except by introspection, and attributing it to others requires a certain degree of faith in unprovable assumptions. The question is: which unprovable assumptions are we willing to make scientifically? Because all human brains are wired in the same way, I am on fairly safe ground in assuming that you have the same basic factory-installed brain functions that I do. While the human brain is similar in many ways to the brains of other mammals, even other vertebrates,[33] it is also different in very significant ways.[34] I thus restrict my discussion of conscious feelings to humans, which makes the problem more manageable. Still, consciousness is a complex and contentious topic that cannot be discussed exhaustively here. I will therefore simply summarize how I believe conscious feelings come about.

I pursued my Ph.D. working with the cognitive neuroscientist Michael Gazzaniga in the late 1970s. Gazzaniga was famous for his work on split-brain patients,[35] in whom the nerve connections between two sides of the brain are surgically cut in an effort to control otherwise intractable epilepsy. Their misfortune has been a source of many important discoveries about how the brain and mind work. I will just mention one set of findings that Gazzaniga and I made that solidified his ideas about consciousness and that sparked my interest in understanding how the non-conscious aspects of the brain work.[36]

Since typically only the left hemisphere of the brain has the capacity for speech, stimuli presented to the right hemisphere of a split-brain patient cannot be talked about. But the right hemisphere can indicate that it saw and perceived a stimulus by using the left hand (corresponding to the right hemisphere) to select a matching picture. For example, in one study we simultaneously showed the patient's left hemisphere a chicken claw and the right hemisphere a snow scene. The patient's left hand then selected a picture of a shovel. When the patient was asked why he made this choice, his left hemisphere (the speaking hemisphere) responded that it saw a chicken and you need a shovel to clean out the chicken shed.[37] The left hemisphere thus used the information it had available to construct a reality that matched the two pieces of information available: it saw a picture of a chicken and it saw its hand selecting a shovel. Given

the patient's rural background, it made sense to him that a shovel and chicken claw go together since a shovel could be used to clean the chicken shed. In other studies, by presenting commands to the right hemisphere, we induced it to wave, stand, or laugh, and asked the left hemisphere, "Why did you do that?" The left hemisphere responded with answers like "I thought I saw a friend out the window so I waved," "I needed a stretch so I stood up," and "You guys are funny."

On the basis of such findings, Gazzaniga developed his theory of consciousness as an interpreter of experience, a means by which we develop a self-story that we use to understand those motivations and actions that arise from non-conscious processes in our brains.[38] In his view, much of what we do in life is controlled by non-conscious processes that we only come to understand by monitoring and interpreting their expression in behavior or in other body states. Since graduate school, I have been trying to understand how systems that operate outside of conscious awareness, such as those that control the expression of defense responses in the presence of threats, work.

When Gazzaniga and I were doing these studies in the 1970s, consciousness research was not in vogue in psychology or neuroscience. The effects of behaviorism were lingering, but in addition, cognitive science had introduced the idea that the mind is basically a non-conscious information processing device.[39] Consciousness can result from this processing, but the underlying non-conscious processing was the main focus of the field.

In the ensuing decades, scientific interest in consciousness skyrocketed. Much progress has been made in pursuing the neural correlates of consciousness, especially by focusing on how the brain creates conscious perceptions of visual stimuli.[40] Most researchers in this field seem to agree that we are not conscious of representations that occur in the primary visual cortex (the part of the visual cortex that first receives stimuli). Some argue that later stages of visual cortex create our conscious visual perceptions and that this is all that is needed for a conscious experience.[41] Others say that while necessary, the visual cortex alone is not sufficient to produce conscious experience of visual phenomena, and that other circuits and functions are required.[42] For example, one argument is that for an individual to be consciously aware of a visual stimulus, the stimulus has to be attended to,[43] which engages additional cortical areas, including the prefrontal cortex and parietal cortex.[44] Attention also allows the raw visual stimulus to be integrated with memory so that the stimulus can be recognized as a particular object, and even an object that may have had certain personal significance in the past. These attention-controlled representations that include objects and memories are often said to occur in a cognitive workspace[45] sometimes called "working memory" (the capacity to hold information in mind temporarily while doing mental work).[46] Different theories propound different ideas about how information that enters working memory ends up being consciously experienced. For example, according to higher-order theories of consciousness, you must have a thought about a stimulus representation in order to be conscious of it (this is in some ways reminiscent of Gazzaniga's interpreter).[47] The global workspace theory of consciousness, on the other hand, says that information has to be broadcast widely from working memory to other areas that then send signals back to the workspace, resulting in further broadcasting and amplification of the signal and thereby creating the conscious perception.[48] A variety of other cognitive theories also emphasize

the importance of attention and working memory in consciousness.[49]

I cast my lot with the general view that emphasizes the role of working memory as a gateway into consciousness, and I remain neutral about what happens next. My goal is not to solve the consciousness problem, but to understand how consciousness – whatever it may be – makes feelings possible. In my view, once information about the presence of a threat is directed to working memory the stage is set for a conscious feeling – an emotion such as fear – to occur. Working memory is not the same thing as consciousness, but in my opinion most of the conscious experiences we have depend on working memory.

So let us pursue the idea that human emotions are conscious experiences that occur when attention directs information about the operation of non-conscious processes to working memory. An important class of non-conscious processes that contribute to feelings are those associated with the activity of what I referred to above as survival functions (functions related to defense, energy management, fluid balance, thermoregulation, and reproduction). The brain circuits that instantiate these functions are survival circuits (certain feelings can also arise without engagement of survival circuits, but I will not focus on those here).[50] But how, concretely, can the operation of the threat/defense survival system trigger the conscious feeling we call fear?

The capacity to detect and respond to threats is an ancient survival mechanism present in all animals, and likely evolutionarily predates both the capacity to be consciously aware of a threat to well-being and the capacity to consciously experience an inner feeling of fear in response to the threat. Circuits that detect and respond to threats in our brains are not fear circuits, not emotion circuits; they do not make feelings. Hard-wired survival circuits are often mistakenly described as emotion circuits (I did this for some time). But these circuits did not evolve to make feelings. They arose, and continue to exist, simply to help animals stay alive and well.

When a threat activates one of these hard-wired circuits, the result is the establishment of a global motivational state in the organism, a condition that spreads throughout the brain and body to mobilize the organism's resources to deal with the danger. Needs and goals that are unrelated to the threat are supplanted by the here-and-now requirements of the situation. The only relevant motivation is self-preservation. The global organismic state that occurs when an organism is in danger can be called a *defensive motive state*.[51] This state includes activity in circuits that control both innate reactions (survival circuits) and goal-oriented actions that help cope with danger.

Motivational states like these not only occur in mammals (monkeys, dogs, cats, rats, bats, whales), but also in other vertebrates (birds, reptiles, fish) and many invertebrates (flies, bees, slugs, worms). All organisms thus have such mechanisms that help them survive in the face of threats. Defensive motive circuit activation greatly influences behavioral and cognitive activities. When a motive state related to danger is active, we become sensitive and hyper-responsive to stimuli associated with danger. The same occurs if a motive state related to food, drink, or sex occurs.

Feeling afraid is an additional factor that can help promote survival, but it is not the most common response in nature. Feeling afraid only occurs in organisms that can be conscious that they are in danger, and I reserve judgment about which organisms other than humans fall into this category. We know humans have conscious feelings but it is far more difficult to know *scientifically* whether other animals do. Thus, the existence of a motive state, and so-called

emotional behavior, is not one and the same as the existence of a conscious feeling. Unless an organism's nervous system has the capacity to consciously experience the motive state, conscious feelings cannot occur.

We know that the human brain can experience emotions in conjunction with motive states. However, all we know scientifically about other animals is that their brain and body respond in certain ways in the presence of stimuli that trigger these motive states. This leads some to argue that we can use behavior to tell us about feelings in animals.[52] But, as previously noted, defensive motive states and corresponding bodily responses can be triggered in humans subliminally and without any feeling;[53] thus, we should not call upon consciousness to explain things in animals that do not require consciousness in humans. Neither, however, should we ignore consciousness entirely. I believe we should address the question of feelings, but in organisms in which we can evaluate them (humans). Again, this is not meant as a denial of animal consciousness, but instead is a decision to deal only with what we can measure scientifically, as opposed to speculating about the implications of those measurements.

One reason why it is so tempting to attribute consciousness to animals is that we have a very strong tendency to interpret the behavior of others in light of how we feel when we act in a certain way. This serves us fairly well in our dealings with most other humans, but begins to cause problems scientifically when we attribute human emotions to infants or animals, since we have no way of verifying what they experience. Consider infants. Subcortical circuits that control innate "emotional" (survival) behaviors develop earlier than cognitive circuits of the cortex. Experts on infant development say that infants can *act* emotional long before they can actually feel emotion.[54] While one could object to this conclusion by saying it is impossible to know what an infant is feeling, that is exactly the point: in the absence of a subject's ability to verbally report (as with infants or animals), it is impossible to know whether he or she is conscious or non-conscious. Ultimately, then, the question of whether animals act but do not feel, or whether they both act and feel, cannot be answered, as we have no direct way of finding out what animals do or do not experience.

Two important questions should be raised about motive states. First, are they causes of defensive behavior or instead are they, like defensive behavior, a consequence of survival circuit activation? The former is the conventional view.[55] My hypothesis, by contrast, is that the motive state is the collective response of the brain to survival circuit activation. Defensive responses thus contribute to defensive motive states rather than the other way around. The second question is whether the motive state itself contributes to conscious feelings by entering working memory, or whether working memory instead only has access to the individual neural components that constitute the motive state. The answer is not known at this point.

Some are concerned that a shift in terminology toward *threat* and *defense*, and thus away from *fear*, will make the work we do on animals less relevant to humans. I think the opposite is the case. By being clear about which processes underlying fear and anxiety involve consciousness and which do not, we greatly expand our ability to elucidate the processes and their relevance to clinical disorders.

People with pathological fear and anxiety suffer from their subjective feelings. If we want to understand the mechanisms underlying the genesis and maintenance of the subjective feelings of fear and anxiety that so trouble these individuals, we

need to understand how implicit (non-conscious) motive states operate in the brain. For example, the fact that people with phobic disorders are hyper-attentive to threat stimuli related to their phobia and exhibit exaggerated responses to such stimuli is easily accounted for in terms of an overactive defensive survival and motivational circuits.[56]

The brain and body consequences of defensive motivational circuit activation generate many (though not all) of the factors that go into a conscious feeling. But the mechanisms of defensive motivational states are not one and the same as the mechanisms that generate feelings. Feelings require more than the presence of a motivational state. That state has to be consciously experienced in order to be consciously felt. This involves the integration of thoughts and long-term declarative memories with defensive state information in working memory.

Even if we never resolve the question of whether other animals have conscious feelings of fear or anxiety, progress on how threats create non-conscious defensive motive states in animals, and how such states can be regulated through drugs or behavioral treatments, could help many people. For one thing, simply turning down the degree of brain and body arousal associated with the motive state will alter the conscious experience of fear or anxiety. But also, successful regulation of the motive state reduces the sensitivity to trigger stimuli and also tones down the heightened reactivity to such triggers that occur in anxiety disorders.

Darwin was right that we have inherited hard-wired circuits from our animal ancestors. These are survival circuits. Their job is to detect significant situations and control behaviors that keep us alive in the face of challenges and that also help us thrive in the presence of opportunities. But Darwin was wrong that we inherited emotional states of mind, such as feelings of fear, from other animals.[57] Survival circuits in subcortical brains are not inherited storehouses of feelings. Feelings are parasitic on the capacity for conscious awareness – which crucially depends on cognitive processes related to attention and working memory – and made possible by cortical circuits.

To understand how the brain makes feelings, consider an analogy to cooking soup.[58] Salt, pepper, garlic, and water are common ingredients in many if not most soups. Put in chicken and it suddenly by definition becomes chicken soup. The amount of salt and pepper can intensify the taste without radically changing the nature of the soup. You can add other ingredients, like celery, turnips, or tomatoes, and still have a variant of chicken soup. Add roux and it becomes gumbo, while curry paste pushes it in a different direction. Substitute shrimp for chicken in any variant and the character again changes. None of these are soup ingredients per se; they are things that exist independent of soup, and that would exist if a soup had never been made. Similarly, emotional feelings emerge from non-emotional ingredients. Specifically, they emerge from the coalescing of non-emotional ingredients in consciousness.[59] The particular ingredients, and the amounts of each, define the character of the feeling. The pot in which feelings cook is working memory.

A defensive motive state provides many of the key ingredients in fear: direct input from the amygdala to cortical areas, brain arousal, body feedback, and initiation of goal-directed behavior.[60] When information about these various activities and their particular neurally encoded characteristics coalesce in working memory via attentional control, together with information about the external stimulus and long-term memories about what that stimulus means, then the resulting feeling that emerges is

some variant of fear. Whether we feel concerned, scared, terrified, alarmed, or panicked depends on the particular characteristics of the internal factors aroused in the brain, factors from the body, and information about the stimulus and its context. In the presence of these neural ingredients, feelings emerge in consciousness similarly to the way the essence of a soup emerges from its ingredients.

Motive states are created from general-purpose mechanisms (such as sensation, memory, arousal, body feedback, and memories and thoughts) but the resulting state is specific to the motivational demands of the moment. A defensive motive state is different from a reproductive (sexual) motive state. And even within a category, the nature of the motive state can vary considerably depending on the circumstances – as can the resultant feeling (for example, consider concern versus fear versus panic).

Emotions resulting from non-conscious motive states emerge in consciousness in a bottom-up fashion, but emotions can also be built from cognitive processes in a top-down fashion without the involvement of motive state ingredients. So-called social emotions are like this (for example, feelings of compassion, pride, and shame). These arise from our assessment of our circumstances.[61] While fear is a prototypical bottom-up emotion, it can also arise from top-down influences. We can think our way into fear and activate a defensive motive state this way. Additionally, we can have intellectual fears, such as the fear of failing in life, of our eventual death, or of alien abduction, that depend on top-down processes rather than simply emerging bottom-up from a motive state as a result of external stimuli.

The enormous complexity of the various conscious manifestations of fear in an individual suggests that there is no one thing that the term *fear* refers to, and there is certainly no "fear module" in the brain that is responsible for all of the states to which we apply the label "fear."[62] Fear, the conscious feeling of being afraid, is what happens when we are aware that certain ingredients have come together to compel a certain interpretation of the state we are in.[63] Anxiety, that sense of worry or apprehension one has when dwelling on the past and/or anticipating the future, is a variation on this theme.

In order to understand feelings like fear, anger, sadness, and joy, we first have to understand how non-conscious, non-emotional ingredients are assembled in consciousness. While consciousness is a hard problem,[64] we do not have to wait on its solution to make progress. There is much to be learned about the non-emotional, non-conscious ingredients that contribute to conscious feelings. Because these are shared to some extent by humans and other animals, we can study the processes across species regardless of whether the species in question have the capacity to be conscious that these states are occurring. The question of whether other animals have feelings is thus reducible to the question of whether they have mechanisms that allow them to be conscious of their own brain states. While we may never answer this question, we have much to learn about human feelings from studies of their non-conscious underpinnings in the brains of humans and animals alike.

ENDNOTES

1 Rene Descartes, *Principia Philosophiae* (Amsterdam: Louis Elzevir, 1644). Available in translation in the public domain at http://www.gutenberg.org/ebooks/4391.

2 Herman Melville, *Moby Dick* (New York: Penguin, 1930).

3 Beth A. Dixon, "Animal Emotion," *Ethics & the Environment* 6 (2001): 22–30.

4 Charles Darwin, *The Expression of the Emotions in Man and Animals* (London: Fontana Press, 1872).

5 The challenges of studying consciousness in animals have been addressed by philosophers and scientists. For some of the many views, see Thomas Nagel, "What Is It Like to Be a Bat?" *Philosophical Review* 83 (1974): 4435–4450; Beth A. Dixon, "Animal Emotion," *Ethics & the Environment* 6 (2001): 22–30; Burrhus F. Skinner, *The Behavior of Organisms: An Experimental Analysis* (New York: Appleton-Century-Crofts, 1938); Donald R. Griffin, "Animal Consciousness," *Neuroscience & Biobehavioral Reviews* 9 (4) (1985): 615–622; Endel Tulving, "Episodic Memory and Autonoesis: Uniquely Human?" in *The Missing Link in Cognition*, ed. Herbert S. Terrace and Janet Metcalfe (New York: Oxford University Press, 2005), 4–56; Nicola S. Clayton and Anthony Dickinson, "Episodic-Like Memory During Cache Recovery by Scrub Jays," *Nature* 395 (6699) (1998): 272–274; Thomas Suddendorf and Michael C. Corballis, "The Evolution of Foresight: What Is Mental Time Travel, and Is It Unique to Humans?" *Behavioral and Brain Sciences* 30 (3) (2007): 299–313, discussion 313–351; Jaak Panksepp, *Affective Neuroscience* (New York: Oxford University Press, 1998); Joseph E. LeDoux, "Emotional Colouration of Consciousness: How Feelings Come About," in *Frontiers of Consciousness: Chichele Lectures*, ed. Lawrence Weiskrantz and Martin Davies (Oxford: Oxford University Press, 2008), 69–130; David B. Edelman and Anil K. Seth, "Animal Consciousness: A Synthetic Approach," *Trends in Neuroscience* 32 (9) (2009): 476–484; Larry Weiskrantz, "The Problem of Animal Consciousness in Relation to Neuropsychology," *Behavioural Brain Research* 71 (1–2) (1995): 171–175; Celia Heyes, "Beast Machines? Questions of Animal Consciousness," in *Frontiers of Consciousness: Chichele Lectures*, ed. Lawrence Weiskrantz and Martin Davies (Oxford: Oxford University Press, 2008), 259–274; J. D. Smith, J. J. Couchman, and M. J. Beran, "The Highs and Lows of Theoretical Interpretation in Animal-Metacognition Research," *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 367 (2012): 1297–1309; J. Metcalfe and A. P. Shimamura, *Metacognition: Knowing about Knowing* (Cambridge, Mass.: Bradford Books, 1994); and H. Terrace and J. Metcalfe, *The Missing Link in Cognition: Origins of Self-Reflective Consciousness* (New York: Oxford University Press, 2004).

6 This essay expands on ideas that I have discussed in recent publications, including LeDoux, "Emotional Colouration of Consciousness: How Feelings Come About"; Joseph E. LeDoux, "Rethinking the Emotional Brain," *Neuron* 73 (4) (2012): 653–676; and Joseph E. LeDoux, "Coming to Terms with Fear," *Proceedings of the National Academy of Sciences* 111 (8) (2014): 2871–2878.

7 Orval H. Mowrer and R. Ross Lamoreaux, "Fear as an Intervening Variable in Avoidance Conditioning," *Journal of Comparative Psychology* 39 (1946): 29–50; Neal E. Miller, "Studies of Fear as an Acquirable Drive: I. Fear as Motivation and Fear-Reduction as Reinforcement in the Learning of New Responses," *Journal of Experimental Psychology* 38 (1948): 89–101; Judson S. Brown and I. E. Farber, "Emotions Conceptualized as Intervening Variables – with Suggestions toward a Theory of Frustration," *Psychological Bulletin* 48 (6) (1951): 465–495; Robert A. Rescorla and Richard L. Solomon, "Two-Process Learning Theory: Relationships between Pavlovian Conditioning and Instrumental Learning," *Psychological Review* 74 (1967): 151–182; Dorothy E. McAllister and Wallace R. McAllister, "Fear Theory and Aversively Motivated Behavior: Some Controversial Issues," in *Fear, Avoidance, and Phobias: A Fundamental Analysis*, ed. M. Ray Denny (Hillsdale, N.J.: Lawrence Erlbaum Associates, 1991), 135–164; Robert C. Bolles and Michael S. Fanselow, "A Perceptual-Defensive-Recuperative Model of Fear and Pain," *Behavioral and Brain Sciences* 3 (1980): 291–323; Fred A. Masterson and Mary Crawford, "The Defense Motivation System: A Theory of Avoidance Behavior," *Behavioral and Brain Sciences* 5 (1982): 661–696; Susan Mineka, "The Role of Fear in Theories of Avoidance Learning, Flooding, and Extinction," *Psychological Bulletin* 86 (1979): 985–1010; Donald J. Levis, "The

Case for a Return to a Two-Factor Theory of Avoidance: The Failure of Non-Fear Interpretations," in *Contemporary Learning Theories: Pavlovian Conditioning and the Status of Traditional Learning Theory*, ed. Stephen B. Klein and Robert R. Mowrer (Hillsdale, N.J.: Lawrence Erlbaum Associates, 1989), 227–277; Bill P. Godsil and Michael S. Fanselow, "Motivation," in *Handbook of Psychology*, ed. Alice F. Healy and Robert W. Proctor (Hoboken, N.J.: John Wiley & Sons, 2013), 32–60; Ralph Adolphs, "The Biology of Fear," *Current Biology* 23 (2) (2013): R79–93; and Jeffrey B. Rosen and Jay Schulkin, "From Normal Fear to Pathological Anxiety," *Psychological Review* 105 (2) (1998): 325–350.

8 Not all those who view emotions as intervening variables eliminate feelings from the discussion. For example, see Adolphs, "The Biology of Fear"; and Rosen and Schulkin, "From Normal Fear to Pathological Anxiety."

9 Paul D. MacLean, "Psychosomatic Disease and the 'Visceral Brain': Recent Developments Bearing on the Papez Theory of Emotion," *Psychosomatic Medicine* 11 (1949): 338–353; Paul D. MacLean, "Some Psychiatric Implications of Physiological Studies on Frontotemporal Portion of Limbic System (Visceral Brain)," *Electroencephalography and Clinical Neurophysiology* 4 (1952): 407–418; Paul D. MacLean, *The Triune Brain in Evolution: Role in Paleocerebral Functions* (New York: Plenum Press, 1990); and Panksepp, *Affective Neuroscience*.

10 See Joseph E. LeDoux, "Cognition and Emotion: Processing Functions and Brain Systems," in *Handbook of Cognitive Neuroscience*, ed. Michael S. Gazzaniga (New York: Plenum Publishing Corp., 1984), 357–368; and Joseph E. LeDoux, "Emotion," in *Handbook of Physiology 1: The Nervous System, Vol. V, Higher Functions of the Brain*, ed. Fred Plum (Bethesda, Md.: American Physiological Society, 1987), 419–460.

11 Antonio Damasio has also distinguished between emotions and feelings. See Antonio Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain* (New York: Gosset/Putnam, 1994).

12 Joseph E. LeDoux, *The Emotional Brain* (New York: Simon and Schuster, 1996); Joseph E. LeDoux, *Synaptic Self: How Our Brains Become Who We Are* (New York: Viking, 2002); and Joshua P. Johansen, Christopher K. Cain, Linnaea E. Ostroff, and Joseph E. LeDoux, "Molecular Mechanisms of Fear Learning and Memory," *Cell* 147 (3) (2011): 509–524.

13 George Mandler and William Kessen, *The Language of Psychology* (New York: John Wiley & Sons, 1959); and Paul M. Churchland, "Folk Psychology and the Explanation of Human Behavior," *Proceedings of the Aristotelian Society* 62 (1988): 209–221.

14 Kurt Danziger, *Naming the Mind: How Psychology Found Its Language* (London: Sage Publications, 1997); and Mandler and Kessen, *The Language of Psychology*.

15 LeDoux, "Rethinking the Emotional Brain."

16 LeDoux, "Coming to Terms with Fear."

17 Joseph E. LeDoux, "Emotion Circuits in the Brain," *Annual Review of Neuroscience* 23 (2000): 155–184; Michael Davis, "Neural Systems Involved in Fear and Anxiety Measured with Fear-Potentiated Startle," *American Psychologist* 61 (8) (2006): 741–756; Michael S. Fanselow and Andrew M. Poulos, "The Neuroscience of Mammalian Associative Learning," *Annual Review of Psychology* 56 (2005): 207–234; and Stephen Maren, "Neurobiology of Pavlovian Fear Conditioning," *Annual Review of Neuroscience* 24 (2001): 897–931.

18 LeDoux, "Coming to Terms with Fear."

19 Robert J. Blanchard and D. Caroline Blanchard, "Crouching as an Index of Fear," *Journal of Comparative and Physiological Psychology* 67 (1969): 370–375; Robert C. Bolles and Michael S. Fanselow, "A Perceptual-Defensive-Recuperative Model of Fear and Pain," *Behavioral and Brain Sciences* 3 (1980): 291–323.

20 Johansen, Cain, Ostroff, and LeDoux, "Molecular Mechanisms of Fear Learning and Memory"; Maren, "Neurobiology of Pavlovian Fear Conditioning"; and Hans-Christian Pape and Denis Pare, "Plastic Synaptic Networks of the Amygdala for the Acquisition, Expression, and Extinction of Conditioned Fear," *Physiology Reviews* 90 (2) (2010): 419–463.

[21] Elizabeth A. Phelps, "Emotion and Cognition: Insights from Studies of the Human Amygdala," *Annual Review of Psychology* 57 (2006): 27 – 53; Christian Buchel and Raymond J. Dolan, "Classical Fear Conditioning in Functional Neuroimaging," *Current Opinion in Neurobiology* 10 (2) (2000): 219 – 223; Susan Mineka and Arne Ohman, "Phobias and Preparedness: The Selective, Automatic, and Encapsulated Nature of Fear," *Biological Psychiatry* 52 (10) (2002): 927 – 937; Patrik Vuilleumier and Gilles Pourtois, "Distributed and Interactive Brain Mechanisms During Emotion Face Perception: Evidence from Functional Neuroimaging," *Neuropsychologia* 45 (1) (2007): 174 – 194; and Paul J. Whalen, Jerome Kagan, Robert G. Cook, F. Caroline Davis, Hackjin Kim, Sara Polis, Donald G. McLaren, Leah H. Somerville, Ashly A. McLean, Jeffrey S. Maxwell, and Tom Johnstone, "Human Amygdala Responsivity to Masked Fearful Eye Whites," *Science* 306 (5704) (2004): 2061.

[22] Joseph LeDoux, "The Amygdala," *Current Biology* 17 (20) (2007): R868 – 874; Johansen, Cain, Ostroff, and LeDoux, "Molecular Mechanisms of Fear Learning and Memory"; and Stephen Maren, "Neurobiology of Pavlovian Fear Conditioning," *Annual Review of Neuroscience* 24 (2001): 897 – 931.

[23] David G. Amaral, "The Amygdala, Social Behavior, and Danger Detection," *Annals of the New York Academy of Sciences* 1000 (2003): 337 – 347; Ned H. Kalin, Steven E. Shelton, and Richard J. Davidson, "The Role of the Central Nucleus of the Amygdala in Mediating Fear and Anxiety in the Primate," *The Journal of Neuroscience* 24 (24) (2004): 5506 – 5515; and Andy M. Kazama, Eric Heuer, Michael Davis, and Jocelyne Bachevalier, "Effects of Neonatal Amygdala Lesions on Fear Learning, Conditioned Inhibition, and Extinction in Adult Macaques," *Behavioral Neuroscience* 126 (3) (2012): 392 – 403.

[24] See Thomas J. Carew, Edgar T. Walters, and Eric R. Kandel, "Associative Learning in *Aplysia*: Cellular Correlates Supporting a Conditioned Fear Hypothesis," *Science* 211 (4481) (1981): 501 – 504; Eric R. Kandel, "The Molecular Biology of Memory: Camp, Pka, Cre, Creb-1, Creb-2, and Cpeb," *Molecular Brain* 5 (2012): 14; David L. Glanzman, "Common Mechanisms of Synaptic Plasticity in Vertebrates and Invertebrates," *Current Biology* 20 (1) (2010): R31 – R36; and Jerry C. Yin and Timothy Tully, "Creb and the Formation of Long-Term Memory," *Current Opinion in Neurobiology* 6 (2) (1996): 264 – 268.

[25] Johansen, Cain, Ostroff, and LeDoux, "Molecular Mechanisms of Fear Learning and Memory."

[26] For examples, see Chris Frith, Richard Perry, and Erik Lumer, "The Neural Correlates of Conscious Experience: An Experimental Framework," *Trends in Cognitive Sciences* 3 (3) (1999): 105 – 114; Lawrence Weiskrantz and Martin Davies, eds., *Frontiers of Consciousness: Chichele Lectures* (Oxford: Oxford University Press, 2008); Stanislas Dehaene and Jean-Pierre Changeux, "Experimental and Theoretical Approaches to Conscious Processing," *Neuron* 70 (2) (2011): 200 – 227; Philip D. Zelazo, Morris Moscovitch, and Evan Thompson, eds. *The Cambridge Handbook of Consciousness* (New York: Cambridge University Press, 2007); and Antonio R. Damasio, "Investigating the Biology of Consciousness," *Philosophical Transactions of the Royal Society B: Biological Sciences* 353 (1377) (1998): 1879 – 1882.

[27] For example, Phelps, "Emotion and Cognition: Insights from Studies of the Human Amygdala"; Mineka and Ohman, "Phobias and Preparedness"; and Whalen et al., "Human Amygdala Responsivity to Masked Fearful Eye Whites."

[28] Vuilleumier and Pourtois, "Distributed and Interactive Brain Mechanisms During Emotion Face Perception"; Phelps, "Emotion and Cognition: Insights from Studies of the Human Amygdala"; Mineka and Ohman, "Phobias and Preparedness"; and Whalen et al., "Human Amygdala Responsivity to Masked Fearful Eye Whites."

[29] Piotr Winkielman, Kent C. Berridge, and Julia L. Wilbarger, "Unconscious Affective Reactions to Masked Happy Versus Angry Faces Influence Consumption Behavior and Judgments of Value," *Personality and Social Psychology Bulletin* 31 (1) (2005): 121 – 135.

[30] LeDoux, "Emotional Colouration of Consciousness: How Feelings Come About"; LeDoux, "Rethinking the Emotional Brain"; and LeDoux, "Coming to Terms with Fear."

31 LeDoux, "Rethinking the Emotional Brain"; Robert M. Macnab and D. E. Koshland, Jr., "The Gradient-Sensing Mechanism in Bacterial Chemotaxis," *Proceedings of the National Academy of Sciences* 69 (3) (1972): 2509 – 2512.

32 LeDoux, "Rethinking the Emotional Brain."

33 Walle J.H. Nauta and Harvey J. Karten, "A General Profile of the Vertebrate Brain, with Sidelights on the Ancestry of Cerebral Cortex," in *The Neurosciences: Second Study Program*, ed. Francis O. Schmitt (New York: The Rockefeller University Press, 1970), 7 – 26.

34 Georg F. Striedter, *Principles of Brain Evolution* (Sunderland, Mass.: Sinauer Associates, 2005); Roger Reep, "Relationship Between Prefrontal and Limbic Cortex: A Comparative Anatomical Review," *Brain, Behavior and Evolution* 25 (1984): 5 – 80; Todd M. Preuss, "Do Rats Have Prefrontal Cortex? The Rose-Woolsey-Akert Program Reconsidered," *Journal of Cognitive Neuroscience* 7 (1995): 1 – 24; Steven P. Wise, "Forward Frontal Fields: Phylogeny and Fundamental Function," *Trends in Neurosciences* 31 (2008): 599 – 608; Eva Braak, "On the Structure of IIIab-Pyramidal Cells in the Human Isocortex. A Golgi and Electron Microscopical Study with Special Emphasis on the Proximal Axon Segment," *Journal für Hirnforschung* 21 (1980): 437 – 442; and Katerina Semendeferi, Kate Teffer, Dan P. Buxhoeveden, Min S. Park, Sebastian Bludau, Katrin Amunts, Katie Travis, and Joseph Buckwalter, "Spatial Organization of Neurons in the Frontal Pole Sets Humans Apart from Great Apes," *Cerebral Cortex* 21 (2011): 1485 – 1497.

35 Michael S. Gazzaniga, *The Bisected Brain* (New York: Appleton-Century-Crofts, 1970).

36 See Michael S. Gazzaniga and Joseph E. LeDoux, *The Integrated Mind* (New York: Plenum, 1978).

37 There is some dispute over whether the patient said "shed" or "shit." The former version has been perpetuated and is what I use here, but in either case the implication is the same.

38 Michael S. Gazzaniga, *Mind Matters* (Cambridge, Mass.: MIT Press, 1988); Michael S. Gazzaniga, *The Social Brain* (New York: Basic Books, 1985); and Michael S. Gazzaniga, *Who's in Charge? Free Will and the Science of the Brain* (New York: Ecco, 2012).

39 Ulric Neisser, *Cognitive Psychology* (New York: Appleton-Century-Crofts, 1967); and Howard E. Gardner, *The Mind's New Science: A History of the Cognitive Revolution* (New York: Basic Books, 1987).

40 Francis Crick and Christof Koch, "A Framework for Consciousness," *Nature Neuroscience* 6 (2) (2003): 119 – 126.

41 Ned Block, "Consciousness, Accessibility, and the Mesh between Psychology and Neuroscience," *Behavioral Brain Science* 30 (5 – 6) (2007): 481 – 499, discussion 499 – 548.

42 Crick and Koch, "A Framework for Consciousness"; David Rosenthal, "Higher-Order Awareness, Misrepresentation and Function," *Philosophical Transactions of the Royal Society B: Biological Sciences* 367 (1594) (2012): 1424 – 1438; Dehaene and Changeux, "Experimental and Theoretical Approaches to Conscious Processing"; and Jesse J. Prinz, *The Conscious Brain: How Attention Engenders Experience* (New York: Oxford University Press, 2012).

43 Daniel Bor and Anil K. Seth, "Consciousness and the Prefrontal Parietal Network: Insights from Attention, Working Memory, and Chunking," *Frontiers in Psychology* 3 (2012): 63; and Prinz, *The Conscious Brain: How Attention Engenders Experience*.

44 Hakwan C. Lau and Richard E. Passingham, "Relative Blindsight in Normal Observers and the Neural Correlate of Visual Consciousness," *Proceedings of the National Academy of Sciences* 103 (49) (2006): 18763 – 18768; Stanislas Dehaene and Lionel Naccache, "Towards a Cognitive Neuroscience of Consciousness: Basic Evidence and a Workspace Framework," *Cognition* 79 (2001): 1 – 37; Stanislas Dehaene, Claire Sergent, and Jean-Pierre Changeux, "A Neuronal Network Model Linking Subjective Reports and Objective Physiological Data During Conscious Perception," *Proceedings of the National Academy of Sciences* 100 (2003): 8520 – 8525; and Crick and Koch, "A Framework for Consciousness."

45 Dehaene and Changeux, "Experimental and Theoretical Approaches to Conscious Processing," *Neuron* 70 (2) (2011): 200 – 227; and Katharine McGovern and Bernard J. Baars, "Cog-

nitive Theories of Consciousness," in *The Cambridge Handbook of Consciousness*, ed. Philip D. Zelazo, Morris Moscovitch, and Evan Thompson (New York: Cambridge University Press, 2007), 177–205.

46 Philip N. Johnson-Laird, *The Computer and the Mind: An Introduction to Cognitive Science* (Cambridge, Mass.: Harvard University Press, 1988); Donald A. Norman and Tim Shallice, "Attention to Action: Willed and Automatic Control of Behavior," in *Consciousness and Self-Regulation*, ed. Richard J. Davison, Gary E. Schwartz, and David Shapiro (New York: Plenum, 1980), 1–18; Timothy Shallice, "Information Processing Models of Consciousness," in *Consciousness in Contemporary Science*, ed. Anthony J. Marcel and E. Bisiach (Oxford: Oxford University Press, 1988), 305–333; Alan Baddeley and Graham Hitch, "Working Memory," in *The Psychology of Learning and Motivation, Vol. 8*, ed. Gordon Bower (New York: Academic Press, 1974), 47–89; Alan Baddeley, "The Episodic Buffer: A New Component of Working Memory?" *Trends in Cognitive Science* 4 (11): 417–423; and Daniel L. Schachter, "Toward a Cognitive Neuropsychology of Awareness: Implicit Knowledge and Anosognosia," *Journal of Clinical and Experimental Neuropsychology* 12 (1990): 155–178.

47 Rosenthal, "Higher-Order Awareness, Misrepresentation and Function."

48 Dehaene and Changeux, "Experimental and Theoretical Approaches to Conscious Processing" and McGovern and Baars, "Cognitive Theories of Consciousness."

49 Philip N. Johnson-Laird, "A Computational Analysis of Consciousness," in *Consciousness in Contemporary Science*, ed. Anthony J. Marcel and E. Bisiach (Oxford: Oxford University Press, 1993), 357–368; Shallice, "Information Processing Models of Consciousness"; Daniel L. Schacter, "Toward a Cognitive Neuropsychology of Awareness: Implicit Knowledge and Anosognosia," *Journal of Clinical and Experimental Neuropsychology* 12 (1) (1990): 155–178; and Prinz, *The Conscious Brain: How Attention Engenders Experience*.

50 LeDoux, "Rethinking the Emotional Brain"; and LeDoux, "Coming to Terms with Fear."

51 Masterson and Crawford, "The Defense Motivation System: A Theory of Avoidance Behavior"; and Bolles and Fanselow, "A Perceptual-Defensive-Recuperative Model of Fear and Pain."

52 For example, see Jaak Panksepp, *Affective Neuroscience* (New York: Oxford University Press, 1998); and Marc Bekoff, *The Emotional Lives of Animals: A Leading Scientist Explores Animal Joy, Sorrow, and Empathy – and Why They Matter* (Novato, Calif.: New World Library, 2007).

53 Phelps, "Emotion and Cognition: Insights from Studies of the Human Amygdala"; Buchel and Dolan, "Classical Fear Conditioning in Functional Neuroimaging"; Mineka and Ohman, "Phobias and Preparedness"; Vuilleumier and Pourtois, "Distributed and Interactive Brain Mechanisms During Emotion Face Perception"; and Whalen et al., "Human Amygdala Responsivity to Masked Fearful Eye Whites."

54 Michael Lewis, *The Rise of Consciousness and the Development of Emotional Life* (New York: Guilford Press, 2013).

55 Clifford T. Morgan, Physiological Psychology (New York: McGraw-Hill, 1943); Dalbir Bindra, "A Unified Interpretation of Emotion and Motivation," *Annals of the New York Academy of Science* 159 (1969): 1071–1083; Wallace R. McAllister and Dorothy E. McAllister, "Behavioral Measurement of Conditioned Fear," in *Aversive Conditioning and Learning*, ed. F. R. Brush (New York: Academic Press), 105–179; Robert A. Rescorla and Richard L. Solomon, "Two-Process Learning Theory: Relationships between Pavlovian Conditioning and Instrumental Learning," *Psychological Review* 74 (1971): 151–182; Miller, "Studies of Fear as an Acquirable Drive"; and Orval H. Mowrer, *Learning Theory and Behavior* (New York: Wiley, 1960).

56 Aaron T. Beck and David A. Clark, "An Information Processing Model of Anxiety: Automatic and Strategic Processes," *Behavior Research and Therapy* 35 (1) (1997): 49–58; and Richard J. McNally, George E. English, and Howard J. Lipke, "Assessment of Intrusive Cognition in PTSD: Use of the Modified Stroop Paradigm," *Journal of Traumatic Stress* 6 (1993): 33–41.

57 LeDoux, "Rethinking the Emotional Brain." See also Lisa F. Barrett, "Are Emotions Natural Kinds?" *Perspectives on Psychological Science* 1 (2006): 28–58; Lisa F. Barrett, Kristen A. Lind-

quist, Eliza Bliss-Moreau, Seth Duncan, Maria Gendron, Jennifer Mize, and Lauren Brennan, "Of Mice and Men: Natural Kinds of Emotions in the Mammalian Brain? A Response to Panksepp and Izard," *Perspectives on Psychological Science* 2 (3) (2007): 297–311.

58 Lisa Barrett has proposed a similar analogy.

59 I've long proposed that feelings are products of non-emotional ingredients (sensory, memory, and "emotional,"or what I now call survival-circuit information). This idea also appears in recent articles by Lisa Barrett and James Russell, who emphasize that emotions are psychological constructions built from non-emotional ingredients. See James A. Russell, "Core Affect and the Psychological Construction of Emotion," *Psychological Review* 110 (1) (2003): 145–172; Barrett, "Are Emotions Natural Kinds?"; and Barrett et al., "Of Mice and Men: Natural Kinds of Emotions in the Mammalian Brain?"

60 Antonio Damasio emphasizes body feedback. See Damasio, *Descartes' Error*; and Antonio Damasio and Gil B. Carvalho, "The Nature of Feelings: Evolutionary and Neurobiological Origins," *Nature Reviews Neuroscience* 14 (2) (2013): 143–152. In my model, feedback is one of the many ingredients that contribute to feelings.

61 There is much literature on the role of cognitive appraisal in emotion. See Klaus R. Scherer, "Emotion as a Multicomponent Process: A Model and Some Cross-Cultural Data," *Review of Personality and Social Psychology* 5 (1984): 37–63; Klaus R. Scherer, Angela Schorr, and Tom Johnstone, eds., *Appraisal Processes in Emotion: Theory, Methods, Research* (London: London University Press, 2001); Nico H. Frijda, "The Place of Appraisal in Emotion," *Cognition and Emotion* 7 (1993): 357–387; and Andrew Ortony, Gerald L. Clore, and Allan Collins, *The Cognitive Structure of Emotions* (Cambridge: Cambridge University Press, 1988).

62 Barrett, "Are Emotions Natural Kinds?"; and Barrett et al., "Of Mice and Men: Natural Kinds of Emotions in the Mammalian Brain?"

63 See again Barrett, "Are Emotions Natural Kinds?"; and Barrett et al., "Of Mice and Men: Natural Kinds of Emotions in the Mammalian Brain?"; as well as James A. Russell, "Core Affect and the Psychological Construction of Emotion," *Psychological Review* 110 (1) (2003): 145–172.

64 David Chalmers, *The Conscious Mind* (New York: Oxford University Press, 1996).

Working Memory Capacity: Limits on the Bandwidth of Cognition

Earl K. Miller & Timothy J. Buschman

Abstract: Why can your brain store a lifetime of experiences but process only a few thoughts at once? In this article we discuss "cognitive capacity" (the number of items that can be held "in mind" simultaneously) and suggest that the limit is inherent to processing based on oscillatory brain rhythms, or "brain waves," which may regulate neural communication. Neurons that "hum" together temporarily "wire" together, allowing the brain to form and re-form networks on the fly, which may explain a hallmark of intelligence and cognition: mental flexibility. But this comes at a cost; only a small number of thoughts can fit into each wave. This explains why you should never talk on a mobile phone when driving.

EARL K. MILLER is the Picower Professor of Neuroscience at the Massachusetts Institute of Technology.

TIMOTHY J. BUSCHMAN is an Assistant Professor in the Department of Psychology at Princeton University.

*(*See endnotes for complete contributor biographies.)*

Working memory holds the contents of our thoughts. It acts as a mental sketchpad, providing a surface on which we can place transitory information to hold it "in mind." We can then "think" by manipulating this information, such as by combining it with other items or transforming it into something new. For example, working memory allows us to remember phone numbers, do mental arithmetic, and plan errands.

Given its fundamental role in thought it is surprising that working memory has such a severely limited capacity: we can only hold a few thoughts in our consciousness at once. In other words, the surface area of our mental sketchpad is quite small. This limitation is obvious whenever we try to multitask, such as when we attempt to talk on the phone while writing an email, and it is why using our mobile phones while driving increases accident risk, even if we are using a hands-free set.

This stands in contrast to other mental abilities that are not limited, such as long-term memory storage. We can store (seemingly) a lifetime of experiences, but, for some reason, we can only consciously express these thoughts a few at a time. This limited capacity may be fundamentally responsible for the

doi:10.1162/DAED_a_00320

cognitive architecture of our brains: researchers believe it to be the reason we have evolved the ability to focus on one thing at a time (to "attend" to something). Despite being well studied, no one has yet confirmed why working memory is limited. In this essay, we will review some of what is known about working memory capacity and offer our theory of why consciousness may have this limit.

Though we may feel that we are able to perceive most of the world around us, this sensation is, in fact, an illusion constructed by our brains. In reality, we sense a very small part of the world at any point in time; we "sip" at the outside world through a straw. Our brain takes these small bits of data and pieces them together to present an impression of a coherent and holistic scene. Examples of this limitation are abundant: consider the puzzles in which you must identify ten differences between two similar pictures. The brain requires a surprisingly long time to accomplish this, despite the two pictures being side by side and the changes often being obvious, such as the total disappearance of a building or tree. This effect is often referred to as *change blindness* and is a regular occurrence of natural vision. (Another example of change blindness is the large number of editing mistakes we fail to notice in movies.)

The limited bandwidth of consciousness is also apparent in studies of working memory capacity. In these experiments, subjects briefly view a screen with a variable number of objects (such as colored squares) and then, after a delay of a few seconds in which they must hold the objects in memory, they are shown another screen of objects, one of which may be different from what was previously shown.[1] Subjects are then asked whether something has changed, and if so, to identify how it has changed (whether it used to be a different color or shape). When the number of objects on-screen increases beyond a few items, subjects begin to make errors (by missing changes), indicating that their working memory capacity has been exceeded. Experiments such as this have revealed that the average adult human can only process and retain four or five objects at a time (similar to the average monkey, as shown below).[2] The exact capacity of the brain varies by individual; some can remember only one or two items and others can remember up to seven.[3] Interestingly, an individual's capacity is highly correlated with measures of fluid intelligence, suggesting that individual capacity limits may be a fundamental restriction on high-level cognition.[4] This seems intuitive: if you can hold more information in mind at the same time, then more ideas can be combined at once into sophisticated thought.

But what is the nature of the capacity limitation? Do we simply miss new items once we have filled our thoughts with four or five? Or do we always try to take in as much information as possible, eventually spreading ourselves too thin when there are more than four or five objects present? In fact, both may be true.

Models of a strict limit on the number of items you can hold in mind posit that this is because working memory has a limited number of discrete "slots," each of which can independently hold information. And once you fill those slots, you can no longer store any new information. In contrast, other models predict that our limited capacity is due to our spreading ourselves too thin. They suggest that working memory is a flexible resource, a neural pool that can be subdivided among objects. You do not stop storing new information after you reach a fixed capacity as in the slot model; rather, as new information is received, the resource pool is continually divided until the information is spread so thin that it can no longer be ac-

curately recalled (and therefore cannot support behavior). Much evidence has been marshaled on behalf of both models, primarily from studies of the patterns of errors humans make on tests of cognitive capacity. Recently, we examined the neurophysiological mechanisms underlying capacity limits in monkeys. We found an intriguing possibility: both the slot and flexible-resource models are correct, albeit for different reasons.

The advantages of animal work include tighter control over gaze as well as more precise measurements of neural activity than is possible with human subjects. These advantages allowed us to dig deeper into the phenomenon and led to a surprising discovery. The monkeys, like humans, had an overall capacity of four objects. But the monkeys' overall capacity was actually composed of two separate capacities of two objects each in the right and left visual hemifields (to the right and left of the center of gaze) that were *independent* of each other. The processing of objects on the right half of gaze is unaffected by objects in the left half of gaze, regardless of how many objects there were on the left (and vice versa). But adding even one object on the same side of gaze as another object resulted in a decrement in performance. It was as if the monkeys had two separate brains, each one assigned to the right or left half of vision. This right/left independence was surprising, though research focusing on a different type of task might have predicted it: humans have independent capacities to track moving objects in the right and left visual hemifields.[5]

This phenomenon is likely related to the fact that the right and left visual hemifields are respectively processed by the left and right cerebral hemispheres. This suggests that the two cerebral hemispheres can operate somewhat independently, at least for the processing required for visual information to reach awareness. Indeed, the apparent split between the two hemispheres recalls some of the initial observations of humans who had their cerebral hemispheres split to control epilepsy. Without careful testing, these subjects usually appeared normal. Thus, there may be something of a split even in the intact brain: the two visual hemifields/cerebral hemispheres act like two independent slots for processing and holding visual information. At first blush, this seems to support the slot model, with slots for both the left and right fields of vision. But we also found evidence to support the flexible-resource model *within* each visual hemifield: on each side of visual space, information was shared and spread among objects. To show this, we looked more closely at how neurons encoded the contents of working memory.

A pure slot model predicts that encoding an object is all-or-none: if the brain successfully remembers an object, there should be an equal amount of information about it regardless of how many other objects are in the array. But we found that even when a given object was successfully encoded and retained, neural information about that specific object was reduced when another object was added to the same visual hemifield, as if a limited amount of neural information was spread between the objects. The slot model also predicts that if a subject misses an object, no information about it should be recorded in the brain; either an object fills a slot, and is remembered, or not. By contrast, the flexible-resource model suggests that even when a subject misses an object, some information about the object could have been recorded in the brain, just not enough to support conscious perception and memory. This latter prediction is exactly what we found: even when a subject did not consciously perceive the object, the brain still recorded a significant, albeit reduced, amount of information.

In sum, the two cerebral hemispheres (visual hemifields) act like discrete resource slots; within them, neural information is divided among objects in a graded fashion. A number of recent studies in humans support such a hybrid model, finding that there are multiple slots that can store graded information about objects.[6] Thus, capacity limits may reflect interplay or blend between different types of underlying constraints on neural processing. On the one hand, neural processing on the right and left halves of visual space can be slot-like, akin to buckets that can hold a maximum volume of water (information). But, on the other hand, within each cerebral hemisphere there is no limit to the number of objects (thoughts) in each bucket. The limitation is inherent to the information, not the number of objects: if there are too many items in the bucket, only a few can get wet enough (have enough information devoted to them) to reach consciousness. The rest may get a little damp, but it is not enough to act upon.

Whether or not the two cerebral hemispheres have independent capacities for information other than vision remains to be determined. It may prove only to be a visual phenomenon, due to the fact that the right and left of gaze are primarily processed in the left and right cerebral hemispheres, respectively. But even if this independence is limited to vision, it has clear practical implications. For example, taking into account the separate capacities of the right and left of gaze can help in the design of heads-up displays, such as on automobile windshields, maximizing the amount of information that drivers can glean in each glance, or providing information without overloading their capacity to fully process important visual scenes, such as the road in front of them.

So far we have seen that despite our impression that we can store and perceive a significant amount of visual information at once, this is not the case. We can only simultaneously think about a very limited amount of information. Our brains knit together these sips of information to give us the illusion that we have a much larger functional bandwidth. (Again, this is something to keep in mind the next time you are driving and have an urge to reach for your mobile phone.) But this still does not explain why there is a capacity limit for conscious thought. What about the brain's functioning dictates such a small bandwidth?

Why can't you hold one thousand thoughts in mind simultaneously, or even just one hundred? There is mounting evidence that the brain uses oscillatory rhythms (brain waves) for communication, especially for processes underlying high-level cognition. The theory is that the brain forms networks by synchronizing oscillations (rhythmic or repetitive neural activity) of the neurons that make up that network. Neurons that "hum" together form networks, and because only so much information can fit into each oscillatory cycle, any communication system based on an oscillating signal will naturally have a limitation on bandwidth. But before delving into the content limits of an oscillatory cycle, what is the evidence supporting a role for oscillatory activity in brain function to begin with?

It has long been known that the brain has large populations of neurons that oscillate in synchrony. These so-called brain waves occur across a wide range of frequencies from very low (less than once a second, or < 1 Hz) to very high (almost once every 15 ms, or > 60 Hz). Brain waves are not random: they vary with mental state. For example, when you are relaxed, your brain tends to show lower frequency waves; but if you suddenly focus on a task, brain regions that are needed to perform that task begin to produce higher frequency waves.

Despite the evidence that brain waves are important for behavior, their exact role in brain function has long been a mystery. Beginning with the pioneering work of physicist and neurobiologist Christoph von der Malsburg, neurophysiologist Wolf Singer, and their colleagues, there has been increasing awareness that synchronizing the oscillations between neurons may be critical in forming functional networks.

Synchronized oscillations are useful for increasing the impact of neural impulses ("spikes," or sharp changes in voltage that neurons use when they signal one another). Spikes from two neurons that arrive simultaneously at a third neuron downstream have a greater impact than if the impulses arrived at different times.[7] Given this, it is easy to imagine how such a mechanism could be useful to focus mental effort on particular representations (when we pay attention). After all, if synchronizing the rhythms of neurons increases the impact of their spikes, then one way to boost the neural signals associated with an attended object would be to increase synchrony between neurons representing it.

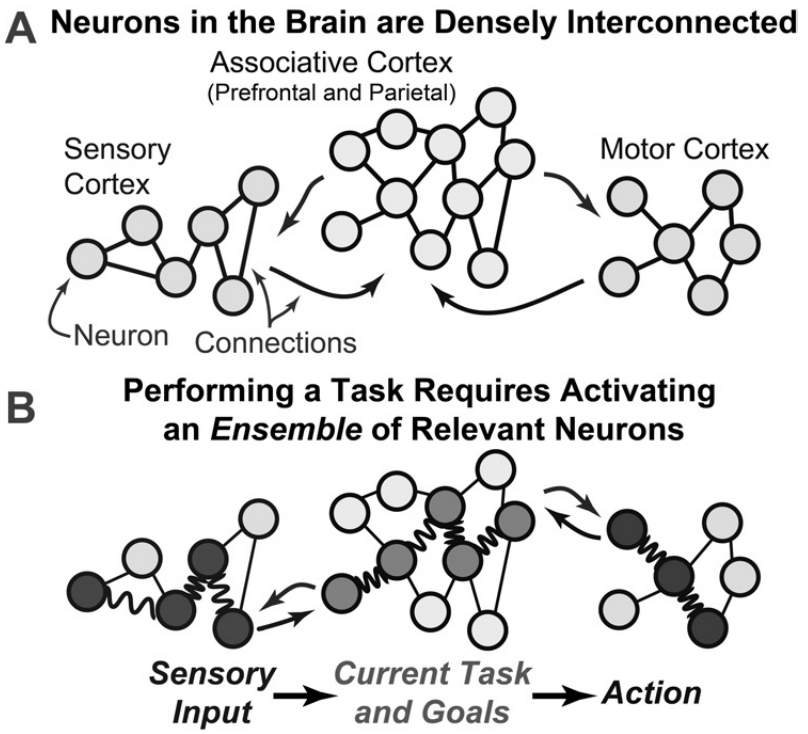
There is growing evidence that this is exactly how attention works. Increased attentional focus increases oscillatory synchrony between the visual cortical neurons that represent the attended stimulus. For example, visual cortical neurons that process a stimulus under attentional focus show increased synchronized gamma band (30 – 90 Hz) oscillations.[8] This higher frequency (> 30 Hz) synchrony may result from interactions within local cortical circuits,[9] the same interactions that underlie the computations of stimulus features.[10] By contrast, sensory cortical neurons representing an unattended stimulus show increased low frequency (< 17 Hz) synchronization. A variety of evidence suggests that low frequencies may help deselect or inhibit the corresponding ensembles (populations of neurons that together underlie a particular thought, perception, memory, or neural computation), perhaps by disrupting the higher frequency.[11]

On a broader scale, synchrony between regions may also regulate communication across brain areas.[12] In short, if two different networks in different brain areas oscillate *in phase* (a particular moment with a neural oscillation, such as a specific "piece" of a brain wave) they are more likely to influence one another because both are in an excited and receptive state at the same time. Conversely, if they are *out of phase*, information will be transmitted poorly. This is supported by observations that interareal oscillatory coherence within and between "cognitive" regions and sensory areas has been found to increase with attention.[13] In other words, if two brain areas are involved in a given cognitive function (such as visual attention), they increase their synchrony during that function.

We have discussed how synchronized rhythms can change the flow of information between neurons and between brain regions. Recent work has begun to suggest that synchrony may not only control communication between networks, it may actually form the networks themselves. The classic model suggests that if neurons are anatomically connected, then they are part of the same network; but it may be that anatomy dictates which neurons are *capable* of forming networks. The actual formation of the networks may instead come through synchrony (Figure 1). In other words, anatomy is like a system of roads; synchrony is the traffic. Importantly, dynamic formation of ensembles by oscillatory synchrony may underlie cognitive flexibility: our ability to rapidly change thoughts and behavior from one moment to the next.

Consider, for example, what is widely assumed to be the basic element of a thought: a group of neurons that are ac-

Figure 1
Synchronous Oscillations Create Neural Ensembles



(A) The human brain consists of almost one hundred billion neurons that form a dense network of connections. This network of neurons and their trillions of connections encapsulate all possible behaviors (and their associated sensations, thoughts, and actions). (B) In order to execute a particular behavior, synchrony activates only those neurons and connections relevant to the current task. Source: Figure prepared by authors.

tive together. Such an ensemble can form a perception, memory, or idea. But how does the brain form a particular neural ensemble for a specific thought? This is not straightforward; there are billions of neurons linked to each other through trillions of connections. This is further complicated because neurons have multiple functions, particularly at "higher," more cognitive levels of the brain.[14] Thus, many neurons inhabit many different ensembles and, conversely, ensembles for different thoughts share some of the same neurons. If anatomy were all there were to forming ensembles, then attempting to activate one ensemble would result in activity that extended to other ensembles, and subsequently a jumble of thoughts.

We propose that the role of synchrony is to dynamically "carve" an ensemble from a greater heterogeneous population of neurons[15] by reinforcing mutual activation between the neurons that form the ensemble.[16] Because ensemble membership would depend on which neurons are oscillating in synchrony at a given moment, ensembles could flexibly form, break-apart, and re-form *without changing their anatomical structure*. In other words, formation of ensembles by rhythmic synchrony endows

thought with flexibility, a hallmark of higher cognition. Humans can quickly adapt and change their thoughts and behaviors in order to tailor them to the constantly changing demands of our complex world. Thus, networks have to be assembled, deconstructed, and reconfigured from moment to moment as our foci, plans, and goals change. This is not meant to downplay the role of neural plasticity in changing the weights of connections between neurons and in forming new anatomical connections; it is always important to build and maintain roads.

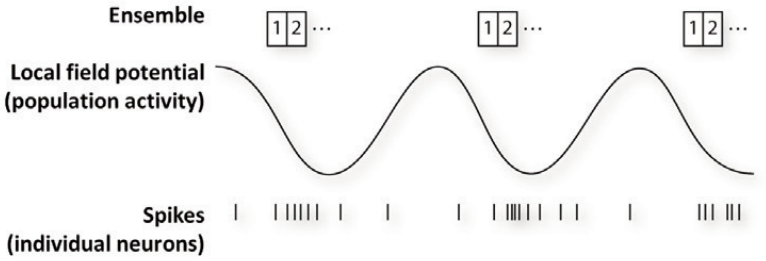
Through a study in which we trained monkeys to switch back and forth between two tasks, we recently found evidence that synchronized oscillations can provide the substrate for dynamic formation of ensembles. As the monkeys switched tasks, different sets of neurons in the prefrontal cortex showed synchronous oscillations – one for each task – in the beta band (about 25 Hz) synchrony, as if the neurons were switching from one network to the other.[17] Importantly, many of the neurons were multifunctional, synchronizing their activity to one ensemble or the other depending on the task at hand. This supports the idea that synchrony can dynamically form (and disassemble) ensembles from anatomical networks of neurons that participate in multiple ensembles.

Interestingly, one of the two tasks was much easier for the monkeys to perform; and when the monkeys prepared to engage in the harder task, the neurons that formed the network for the *easier* task showed synchrony in a low-frequency alpha range (about 10 Hz). Alpha waves have been associated with suppression or inattention to a stimulus[18] and are therefore thought to inhibit irrelevant processes.[19] In our experiment, alpha oscillation inhibition seemed to be acting to quiet the dominant network (the one needed for the easier task), which would have interfered with the network needed for the current, more challenging task. This suggests that synchronous oscillations helped control the formation of ensembles.[20] Higher (beta) frequencies defined the two task networks while lower (alpha) frequencies were used to somehow disrupt formation of the stronger network (and thus prevent an erroneous reflexive reaction) when the weaker network had to be used.

If synchronized rhythms form neural ensembles, it follows to wonder how it is that the brain can form more than one ensemble at a time. After all, would not two rhythmically defined ensembles inadvertently synchronize to each other, merging together and distorting the information they represent? In response, some researchers have proposed that the brain forms more than one ensemble at a time by oscillating different ensembles slightly out of phase with one another.

According to this theory, neurons that are part of a specific ensemble do not only synchronize their activity, but they do so by aligning their spikes to specific phases of neuronal population oscillations.[21] By separating thoughts into different phases of population oscillations, our brain can hold multiple thoughts in mind simultaneously (Figure 2).[22] In other words, the brain prevents ensembles from interfering with one another by juggling them, rhythmically activating each in turn (out of phase from each other). We recently reported evidence for this multiplexing when information is held in mind.[23] When monkeys hold multiple objects in working memory, prefrontal neurons encode information about each object at different phases of an ongoing (~ 32 Hz) oscillation. Significantly, there were bumps of information at different phases, yet in all phases the neurons still carried at least some information, supporting a hybrid slot/flexible-resource model. The bumps of information are

Figure 2
Phase Coding



This figure illustrates oscillatory phase-coding. Neural ensembles of the two simultaneous thoughts (thoughts 1 and 2) oscillate at similar frequencies but different phases of the oscillation. In other words, the ensembles line up on different parts of the brain wave. This may explain the severely limited capacity of consciousness; in this model, only a few thoughts can fit in each wave. Source: Earl K. Miller and Timothy J. Buschman, "Brain Rhythms for Cognition and Consciousness," in *Neurosciences and the Human Person: New Perspectives on Human Activities*, ed. Antonio M. Battro, Stanislas Dehaene, and Wolf Joachim Singer (Vatican City: Pontifical Academy of Sciences, Scripta Varia 121, 2013).

somewhat slot-like in the sense that they are specific to certain phases of the oscillation; but they are not strict slots because the bump is a relative increase over information in other phases. The effect was *not* all-or-none, information-here-but-not-there, as is predicted by a strict slot model.

This finally leads us to an explanation for the severe limitation of conscious thought. Phase-based coding has an inherent capacity limitation. You have to fit all the information needed for conscious thought within an oscillatory cycle. Consciousness may thus be a mental juggling act, and only a few balls can be juggled at once. Crucial tests of this hypothesis still need to be conducted, but these findings and theories collectively suggest that bringing thoughts to consciousness may depend on generation of oscillatory rhythms and the precise temporal relationships between them and the spiking of individual neurons.

ENDNOTES

* Contributor Biographies: EARL K. MILLER is the Picower Professor of Neuroscience at the Massachusetts Institute of Technology. His recent publications include articles in such journals as *Science*, *Nature*, *Proceedings of the National Academy of Sciences*, and *Neuron*.

TIMOTHY J. BUSCHMAN is an Assistant Professor in the Department of Psychology at Princeton University. He has published articles in such journals as *Neuron*, *Proceedings of the National Academy of Sciences*, and *Science*.

1 Steven J. Luck and Edward K. Vogel, "The Capacity of Visual Working Memory for Features and Conjunctions," *Nature* 390 (1997): 279 – 281.

2 Nelson Cowan, "The Magical Number 4 in Short-Term Memory: A Reconsideration of Mental Storage Capacity," *Behavioral and Brain Sciences* 24 (1) (2001): 87 – 114.

3 Edward K. Vogel and Maro G. Machizawa, "Neural Activity Predicts Individual Differences in Visual Working Memory Capacity," *Nature* 428 (2004): 748 – 751; and Edward K. Vogel, Andrew W. McCollough, and Maro G. Machizawa, "Neural Measures Reveal Individual Differences in Controlling Access to Working Memory," *Nature* 438 (2005): 500 – 503.

4 Randall W. Engle, Stephen W. Tuholski, James E. Laughlin, and Andrew R. A. Conway, "Working Memory, Short-Term Memory, and General Fluid Intelligence: A Latent-Variable Approach," *Journal of Experimental Psychology: General* 128 (3) (1999): 309 – 331; and Keisuke Fukuda, Edward Vogel, Ulrich Mayr, and Edward Awh, "Quantity, Not Quality: The Relationship between Fluid Intelligence and Working Memory Capacity," *Psychonomic Bulletin & Review* 17 (5) (2010): 673 – 679.

5 George A. Alvarez and Patrick Cavanagh, "Independent Resources for Attentional Tracking in the Left and Right Visual Hemifields," *Psychological Science* 16 (8) (2005): 637 – 643.

6 David E. Anderson, Edward K. Vogel, and Edward Awh, "Precision in Visual Working Memory Reaches a Stable Plateau When Individual Item Limits Are Exceeded," *The Journal of Neuroscience* 31 (3) (2011): 1128 – 1138; and Keisuke Fukuda, Edward Awh, and Edward K. Vogel, "Discrete Capacity Limits in Visual Working Memory," *Current Opinion in Neurobiology* 20 (2) (2010): 177 – 182.

7 A. M. H. J. Aertsen, G. L. Gerstein, M. K. Habib, and G. Palm (with the collaboration of P. Gochin and J. Krüger), "Dynamics of Neuronal Firing Correlation: Modulation of 'Effective Connectivity,'" *Journal of Neurophysiology* 61 (5) (1989): 900 – 917; Rony Azouz and Charles M. Gray, "Dynamic Spike Threshold Reveals a Mechanism for Synaptic Coincidence Detection in Cortical Neurons *in vivo*," *Proceedings of the National Academy of Sciences* 97 (14) (2000): 8110 – 8115; Emilio Salinas and Terrence J. Sejnowski, "Impact of Correlated Synaptic Input on Output Firing Rate and Variability in Simple Neuronal Models," *The Journal of Neuroscience* 20 (16) (2000): 6193 – 6209; Markus Siegel and Peter König, "A Functional Gamma-Band Defined by Stimulus-Dependent Synchronization in Area 18 of Awake Behaving Cats," *The Journal of Neuroscience* 23 (10) (2003): 4251 – 4260; and P. H. E. Tiesinga, J.-M. Fellous, J.V. José, and T. J. Sejnowski, "Information Transfer in Entrained Cortical Neurons," *Network: Computation in Neural Systems* (13) (2002): 41 – 66.

8 Pascal Fries, John H. Reynolds, Alan E. Rorie, and Robert Desimone, "Modulation of Oscillatory Neuronal Synchronization by Selective Visual Attention," *Science* 291 (5508) (2001): 1560.

9 Christoph Börgers, Steven Epstein, and Nancy J. Kopell, "Gamma Oscillations Mediate Stimulus Competition and Attentional Selection in a Cortical Network Model," *Proceedings of the National Academy of Sciences* 105 (46) (2008): 18023 – 18028; and Jessica A. Cardin, Marie Carlén, Konstantinos Meletis, Ulf Knoblich, Feng Zhang, Karl Deisseroth, Li-Huei Tsai, and Christopher I. Moore, "Driving Fast-Spiking Cells Induces Gamma Rhythm and Controls Sensory Responses," *Nature* 459 (2009): 663 – 667.

10 Seung-Hee Lee, Alex C. Kwan, Siyu Zhang, Victoria Phoumthipphavong, John G. Flannery, Sotiris C. Masmanidis, Hiroki Taniguchi, Z. Josh Huang, Feng Zhang, Edward S. Boyden, Karl Deisseroth, and Yang Dan, "Activation of Specific Interneurons Improves V1 Feature Selectivity and Visual Perception," *Nature* 488 (2012): 379 – 383; John H. Reynolds and David J. Heeger, "The Normalization Model of Attention," *Neuron* 61 (2009): 168 – 185; Nathan R. Wilson, Caroline A. Runyan, Forea L. Wang, and Mriganka Sur, "Division and Subtraction by Distinct Cortical Inhibitory Networks *In Vivo*," *Nature* 488 (2012): 343 – 348.

11 Timothy J. Buschman, Eric L. Denovellis, Cinira Diogo, Daniel Bullock, and Earl K. Miller, "Synchronous Oscillatory Neural Ensembles for Rules in the Prefrontal Cortex," *Neuron* 76 (4) (2012): 838 – 846; Satu Palva and J. Matias Palva, "Functional Roles of Alpha-Band Phase Synchronization in Local and Large-Scale Cortical Network," *Frontiers in Psychology* 2 (2011): 204; William J. Ray and Harry W. Cole, "EEG Alpha Activity Reflects Attentional Demands, and Beta Activity Reflects Emotional and Cognitive Processes," *Science* 228 (1985): 750; and Sujith Vijayan and Nancy J. Kopell, "Thalamic Model of Awake Alpha Oscillations and Implications for Stimulus Processing," *Proceedings of the National Academy of Sciences* 109 (45) (2012): 18553 – 18558.

[12] Steven L. Bressler, "Interareal Synchronization in the Visual Cortex," *Behavioural Brain Research* 76 (1996): 37 – 49; A. K. Engel, P. Fries, and W. Singer, "Dynamic Predictions: Oscillations and Synchrony in Top-Down Processing," *Nature Reviews Neuroscience* 2 (2001): 704 – 716; Pascal Fries, "A Mechanism for Cognitive Dynamics: Neuronal Communication through Neuronal Coherence," *Trends in Cognitive Sciences* 9 (10) (2005): 474 – 480; and Salinas and Sejnowski, "Impact of Correlated Synaptic Input on Output Firing Rate and Variability in Simple Neuronal Models."

[13] Conrado A. Bosman, Jan-Mathijs Schoffelen, Nicolas Brunet, Robert Oostenveld, Andre M. Bastos, Thilo Womelsdorf, Birthe Rubehn, Thomas Stieglitz, Peter De Weerd, and Pascal Fries, "Attentional Stimulus Selection through Selective Synchronization between Monkey Visual Areas," *Neuron* 75 (5) (2012): 875 – 888; Timothy J. Buschman and Earl K. Miller, "Top-Down Versus Bottom-Up Control of Attention in the Prefrontal and Posterior Parietal Cortices," *Science* 315 (5820) (2007): 1860 – 1862; Georgia G. Gregoriou, Stephen J. Gotts, Huihui Zhou, and Robert Desimone, "High-Frequency, Long-Range Coupling Between Prefrontal and Visual Cortex During Attention," *Science* 324 (5931) (2009): 1207 – 1210; Yuri B. Saalmann, Ivan N. Pigarev, and Trichur R. Vidyasagar, "Neural Mechanisms of Visual Attention: How Top-Down Feedback Highlights Relevant Locations," *Science* 316 (5831) (2007): 1612 – 1615; and Markus Siegel, Tobias H. Donner, Robert Oostenveld, Pascal Fries, and Andreas K. Engel, "Neuronal Synchronization along the Dorsal Visual Pathway Reflects the Focus of Spatial Attention," *Neuron* 60 (4) (2008): 709 – 719.

[14] Jason A. Cromer, Jefferson E. Roy, and Earl K. Miller, "Representation of Multiple, Independent Categories in the Primate Prefrontal Cortex," *Neuron* 66 (5) (2010): 796 – 807; and Mattia Rigotti, Omri Barak, Melissa R. Warden, Xiao-Jing Wang, Nathaniel D. Daw, Earl K. Miller, and Stefano Fusi, "The Importance of Mixed Selectivity in Complex Cognitive Tasks," *Nature* 497 (2013): 585 – 590.

[15] Thomas Akam and Dimitri M. Kullmann, "Oscillations and Filtering Networks Support Flexible Routing of Information," *Neuron* 67 (2) (2010): 308 – 320.

[16] Thilo Womelsdorf, Jan-Mathijs Schoffelen, Robert Oostenveld, Wolf Singer, Robert Desimone, Andreas K. Engel, and Pascal Fries, "Modulation of Neuronal Interactions through Neuronal Synchronization," *Science* 316 (5831) (2007): 1609 – 1612.

[17] Buschman et al., "Synchronous Oscillatory Neural Ensembles for Rules in the Prefrontal Cortex."

[18] Ian C. Gould, Matthew F. Rushworth, and Anna C. Nobre, "Indexing the Graded Allocation of Visuospatial Attention using Anticipatory Alpha Oscillations," *Journal of Neurophysiology* 105 (3) (2011): 1318 – 1326; and Saskia Haegens, Verónica Nácher, Rogelio Luna, Ranulfo Romo, and Ole Jensen, "α-Oscillations in the Monkey Sensorimotor Network Influence Discrimination Performance by Rhythmical Inhibition of Neuronal Spiking," *Proceedings of the National Academy of Sciences* 108 (48) (2011): 19377 – 19382.

[19] Wolfgang Klimesch, Paul Sauseng, and Simon Hanslmayr, "EEG Alpha Oscillations: The Inhibition-Timing Hypothesis," *Brain Research Reviews* 53 (1) (2007): 63 – 88; and Kyle E. Mathewson, Alejandro Lleras, Diane M. Beck, Monica Fabiani, Tony Ro, and Gabrielle Gratton, "Pulsed Out of Awareness: EEG Alpha Oscillations Represent a Pulsed-Inhibition of Ongoing Cortical Processing," *Frontiers in Psychology* 2 (2011): 99.

[20] N. Kopell, M. A. Whittington, and M. A. Kramer, "Neuronal Assembly Dynamics in the Beta1 Frequency Range Permits Short-Term Memory," *Proceedings of the National Academy of Sciences* 108 (9) (2010): 3779 – 3784.

[21] Pascal Fries, Danko Nikoli, and Wolf Singer, "The Gamma Cycle," *Trends in Neuroscience* 30 (7) (2007): 309 – 316; John J. Hopfield and Andreas V. M. Herz, "Rapid Local Synchronization of Action Potentials: Toward Computation with Coupled Integrate-and-Fire Neurons," *Proceedings of the National Academy of Sciences* 92 (1995): 6655 – 6662; Peter König and Andreas K. Engel, "Correlated Firing in Sensory-Motor Systems," *Current Opinion in Neurobiology* 5 (4) (1995): 511 – 519; Gilles Laurent, "Olfactory Network Dynamics and the Coding of Multi-

dimensional Signals," *Nature Reviews Neuroscience* 3 (2002): 884–895; M. R. Mehta, A. K. Lee, and M. A. Wilson, "Role of Experience and Oscillations in Transforming a Rate Code into a Temporal Code," *Nature* 417 (2002): 741–746; and John O'Keefe and Michael L. Recce, "Phase Relationship between Hippocampal Place Units and the EEG Theta Rhythm," *Hippocampus* 3 (3) (1993): 317–330.

22 Ole Jensen and John E. Lisman, "Hippocampal Sequence-Encoding Driven by a Cortical Multi-Item Working Memory Buffer," *Trends in Neurosciences* 28 (2) (2005): 67–72; and John E. Lisman and Marco A. P. Idiart, "Storage of 7 +/- 2 Short-Term Memories in Oscillatory Subcycles," *Science* 267 (5203) (1995): 1512–1515.

23 Markus Siegel, Melissa R. Warden, and Earl K. Miller, "Phase-Dependent Neuronal Coding of Objects in Short-Term Memory," *Proceedings of the National Academy of Sciences* 106 (50) (2009): 21341–21346.

Consciousness

Terrence J. Sejnowski

Abstract: No one did more to draw neuroscientists' attention to the problem of consciousness in the twentieth century than Francis Crick, who may be better known as the co-discoverer (with James Watson) of the structure of DNA. Crick focused his research on visual awareness and based his analysis on the progress made over the last fifty years in uncovering the neural mechanisms underlying visual perception. Because much of what happens in our brains occurs below the level of consciousness and many of our intuitions about unconscious processing are misleading, consciousness remains an elusive problem. In the end, when all of the brain mechanisms that underlie consciousness have been identified, will we still be asking: "What is consciousness?" Or will the question shift, just as the question "What is life?" is no longer the same as it was before Francis Crick?

TERRENCE J. SEJNOWSKI, a Fellow of the American Academy since 2013, is the Francis Crick Professor at the Salk Institute for Biological Studies and an Investigator at the Howard Hughes Medical Institute. He is also Professor of Biological Sciences at the University of California, San Diego. His primary research interest is computational neuroscience. He is the author of *Liars, Lovers and Heroes: What the New Brain Science Has Revealed About How We Become Who We Are* (with Steven R. Quartz, 2002), *Thalamocortical Assemblies: How Ion Channels, Single Neurons and Large-Scale Networks Organize Sleep Oscillations* (with Alain Destexhe, 2001), and *The Computational Brain* (with Patricia S. Churchland, 1992).

Francis Crick was once asked by his mother what scientific problems he wanted to pursue in life.[1] The young Francis replied that there were only two problems that interested him: the mystery of life and the mystery of consciousness.[2] Crick clearly had a keen sense for what is important, but may not have appreciated the difficulty of these problems. Little did his mother know that, in 1953, her son and James Watson would famously discover the structure of DNA, the loose thread that would eventually unravel one of life's great mysteries. However, Crick was not content with this achievement.

The Salk Institute for Biological Studies was founded in La Jolla, California, in 1960 and Crick was one of the earliest non-resident fellows, a position that entailed an annual visit to help the faculty make important decisions on promotions and new research directions. In 1977, Crick permanently moved to the Salk Institute, partly to shift his research focus to neuroscience – which he believed would have been difficult to do at the Laboratory of Molecular Biology in Cambridge, England – and partly to circumvent the age limit that would have required him to retire from Cambridge University.[3] At the Salk Institute, Crick took up his long-standing interest in consciousness and decided to focus on the question of visual aware-

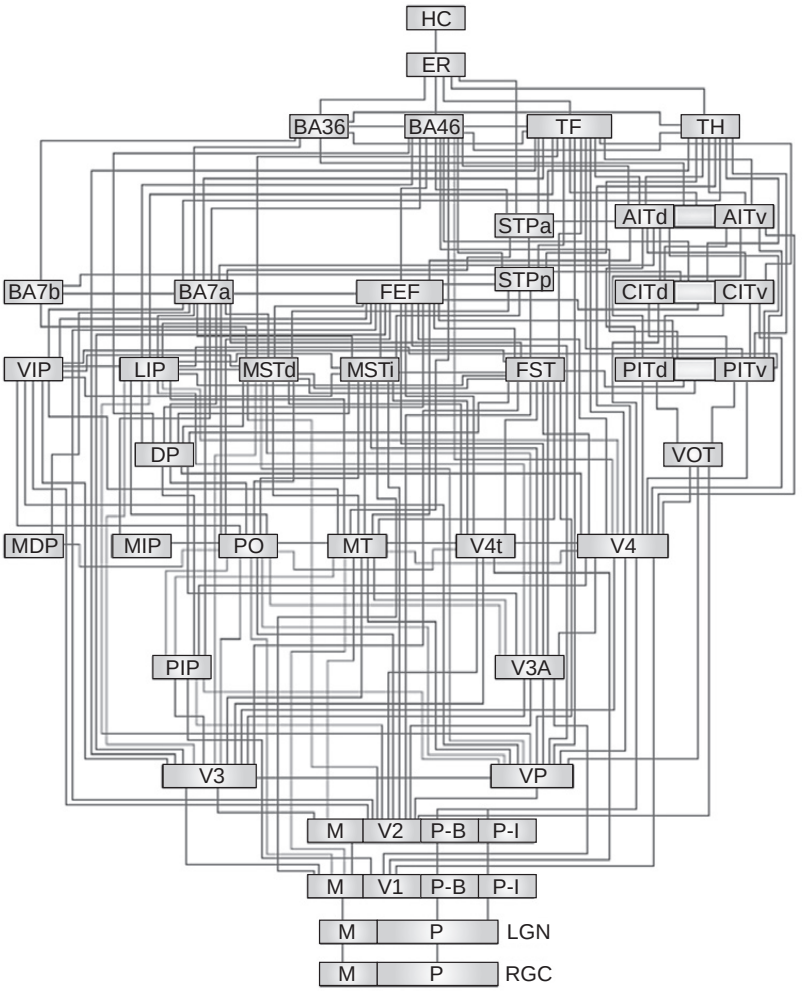
doi:10.1162/DAED_a_00321

ness, since a great deal was already known about the visual parts of the brain and understanding the neural basis of perception would serve as a solid foundation for exploring the neural basis of other aspects of consciousness. This also sidestepped the vagueness of the term *consciousness*, which is used to describe many different phenomena. Together with physicist Gordon Shaw at the University of California, Irvine and neuroscientist V. S. Ramachandran at the University of California, San Diego, Crick founded the Helmholtz Club, a small group of researchers in Southern California who met once a month to discuss problems in vision.[4] In addition, Crick had a steady stream of visitors, including neuroscientist David Marr from the Massachusetts Institute of Technology and physicist Graeme Mitchison from Cambridge University. When I moved to the Salk Institute in 1989, I became the secretary of the Helmholtz Club and helped organize its meetings.

The study of consciousness was out of fashion among biologists in the 1980s, but this did not deter Crick. Visual perception was filled with illusions and mysteries that defied understanding, and he sought explanations for them in anatomical and physiological mechanisms. For example, with Graeme Mitchison, he developed the novel "spotlight of attention" hypothesis. It was well-established that ganglion cells in the eye – neurons in the retina that encode patterns of light on the retina into patterns of spikes – project down the optic nerve to the thalamus (the bilateral brain regions that relay sensory information to the cerebral nerve), which in turn relays the spikes to the visual cortex (Figure 1). But why couldn't the ganglion cells project directly to the cortex? Crick and Mitchison pointed out that there was a feedback projection from the cortex back to the thalamus that, like a spotlight, might highlight parts of the images for further processing.

Crick's closest colleague on the quest for consciousness was neuroscientist Christof Koch, then at the California Institute of Technology, with whom he published a series of papers that explored the neural correlates of consciousness (NCC; the brain structures and neural activities responsible for generating states of conscious awareness).[5] In the case of visual awareness, this meant finding correlations between the firing properties of neurons in different parts of the brain and visual perception. One of their ideas was that we are not aware of what happens in the primary visual cortex, which is the first area of the cerebral cortex[6] to receive input from the retina; rather, they hypothesized, we are only aware of the results of processing at the highest levels of the hierarchy of visual areas in the cortex (Figure 1). Support for this possibility comes from the study of binocular rivalry, in which two different patterns are presented to the two eyes: rather than seeing a blend of the two images, the visual perception flips abruptly between them every few seconds. Neurons in the primary visual cortex respond to both patterns, regardless of which is being consciously perceived at any moment. In the higher levels of the visual hierarchy, however, a larger fraction of the neurons respond only to the perceived image. Thus, it is not enough for a neuron to be firing for it to be a neural correlate of perception. Apparently you are only aware of what is represented in a subset of the active neurons distributed over the hierarchy of visual areas working together in a coordinated way.

In 2004 an epilepsy patient at the UCLA Medical Center whose brain was being monitored to detect the origin of the seizures was shown a series of pictures of celebrities. Electrodes implanted into the memory centers of the patient's brain reported spikes in response to the photos. In one of these patients, a single neuron re-



Visual information from retinal ganglion cells (RGC) in the retina project to the lateral geniculate nucleus (LGN) of the thalamus, whose relay cells project to the primary visual cortex (V1). The hierarchy of cortical areas terminates in the hippocampus (HC). Nearly all of the 187 links in the diagram are bidirectional, with feedforward connection from a lower area and feedback connection from the higher area. Source: Image courtesy of Henry Kennedy; based on Daniel J. Felleman and David C. Van Essen, "Distributed Hierarchical Processing in Primate Visual Cortex," *Cerebral Cortex* 1 (1991): 1 – 47.

sponded vigorously to several pictures of Jennifer Aniston, but not to other famous people.[7] A neuron in another patient would only respond to pictures of Halle Berry, and even to her name, but not to pictures of Bill Clinton or Julia Roberts or the names of other famous people.

Such cells had been predicted fifty years ago when it first became possible to record from single neurons in the brains of cats and monkeys. Researchers thought that in the hierarchy of visual areas of the cerebral cortex, the response properties of the neurons became more and more specific

the higher the neuron was in the hierarchy, perhaps so specific that a single neuron at the top of the hierarchy would only respond to pictures of a single person. This came to be called the "grandmother cell" hypothesis, after the putative neuron in your brain that "recognizes" your grandmother. A team at UCLA led by Itzhak Fried and Christof Koch seemed to have found such cells. Single neurons were also found that recognized specific objects and buildings, like the Sydney Opera House.

Even more dramatic were experiments in which patients looked at a blend of two images representing familiar individuals and were asked to imagine one individual at the expense of the other competing one, while recordings were made from the neurons that preferred one or the other image. The subjects were able to increase the firing rates of the neuron that represented the face they favored in the blend, while simultaneously decreasing the rates of other neurons that preferred the competing face, even though the visual stimulus was not changing. The experimenters then closed the loop by controlling the ratio of the two images in the mixture according to the firing rates of the neurons preferring the images, so the subject could control the input – the ratio of the two faces – by imagining one or the other image. This illustrates that the process of recognition is not simply a passive process, but depends on active engagement of memory and internal attentional control.

Despite this striking evidence, the grandmother cell hypothesis is unlikely to be correct. According to the hypothesis, you perceive your grandmother when the cell is active, so it should not fire to any other stimulus. Only a few hundred pictures were tested, so we really do not know how selective the Jennifer Aniston cell was. Second, the likelihood that the electrode happened to record from the only Jennifer Aniston neuron in the brain is low; it is more likely that there are many thousands of these cells. There must also be many copies of the Halle Berry neuron, and many more for everyone you know and every object you can recognize. Although there are billions of neurons in your brain, you will run out if you try to exclusively represent every object and name that you know by a dedicated population of neurons. Finally, the function of a sensory neuron is only partially determined by its response to sensory inputs. Equally important is the output of the neuron and its downstream impact on behavior.

We are beginning to collect recordings from hundreds of cells simultaneously in mice, monkeys, and humans; and these are leading to a different theory for how neurons collectively perceive and decide.[8] In recordings from monkeys, stimuli and task-dependent signals are broadly distributed over large populations of neurons, each tuned to a different combination of features of the stimuli and task detail.[9] By 2025, it will be possible to record from millions of neurons and to manipulate their firing rates; in addition, new techniques are being developed to distinguish different types of neurons and how they are connected with one another.[10] This could lead to theories beyond the grandmother cell and a deeper understanding of how activity in populations of neurons gives rise to thoughts, emotions, plans, and decisions.

The properties of such distributed representations were first studied in artificial neural networks in the 1980s. Populations of simple model neurons called "hidden units" were trained to map between a set of input units and output units; these hidden units developed patterns of activity for each input that was highly distributed and similar to the variety that has been observed in populations of cortical neurons.[11] For example, the input units might represent faces from many different

angles and the output units might represent the names of the people. After being trained on many examples, each of the hidden units of neurons coded a different combination of features of the input units, such as fragments of eyes, noses, or head shapes, which helped to distinguish between different individuals.

A distributed representation can be used to recognize many versions of the same object, and the same set of neurons can recognize many different objects by differentially weighting their outputs. Moreover, the network can extrapolate general rules from the examples, allowing it to correctly classify new inputs that were not a part of the training set (a process called generalization). Much more powerful versions of these early neural network models, which have over twelve layers of hidden units in a hierarchy like that of our visual cortex (Figure 1) and which use "deep learning" to adjust billions of synaptic weights (strength of influence the firing of one neuron has on another neuron), are now able to recognize tens of thousands of objects in images. When individual hidden units are tested in the same way neurophysiologists record from neurons in the visual cortex, sometimes one simulated neuron near the top of the hierarchy is found to develop a specific preference for one of the objects. However, the performance of the neural network does not appreciably change when such a unit is cut out of it, since the remaining neurons carry redundant signals representing the object. The robustness of the performance of networks against damage is a major difference between the architecture of the brain and that of digital computers.

How many neurons are needed to discriminate between many similar objects such as faces? From imaging studies we know that many areas of the brain respond to faces, some with a high degree of selectivity. To answer this question, we would need to sample many neurons widely from these areas. There are also sound theoretical arguments suggesting minimal numbers of neurons in the representation of an object. First, sparse coding would be more energy-efficient. Second, learning a new object in the same population of neurons interferes with the others being represented in the same population. An effective and efficient representation would be sparsely distributed; that is, it would involve a relatively small fraction of all the neurons, but these would be widely distributed throughout the brain.

Another aspect of visual awareness is the brain's efforts to register events, such as flashes of light, as occurring at specific times. The time delay of neurons in the visual cortex in response to a flashed visual stimulus varies from 25 to 100 milliseconds (ms), often within the same region of the cortex. Nonetheless, we can determine the order of two flashes that occur within 40 ms of each other, and the order of two sounds with less than a 10 ms time difference. To make this even more paradoxical, the processing in the retina itself takes a certain amount of time, which is not fixed but depends on intensity of the flash, so that there is a difference in the arrival time of the first spike from a dim and a bright flash, even though they appear to occur simultaneously. This raises the question of why perceptions seem to have a unity that is not at all apparent from the temporally and spatially distributed patterns of activity throughout the cortex.

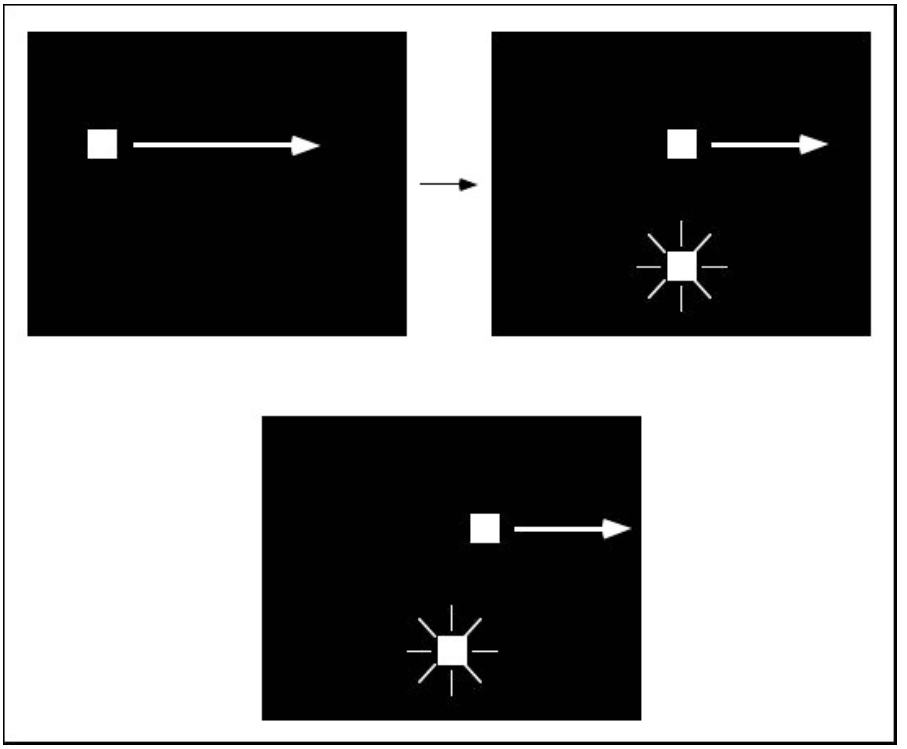
The question of simultaneity becomes even more vexing when we make cross-modal comparisons. As you are watching someone chop down a tree, you simultaneously see and hear the ax hit the tree, even though the speed of sound is much less than that of light. Moreover, the illusion of simultaneity is maintained as the distance from the tree increases,[12] even

though the absolute delay between the visual and auditory signals as they reach your brain can vary over 80 ms before the illusion is broken and the sound is no longer simultaneous with the ax hit.

Researchers who study the temporal aspects of vision have uncovered another phenomenon called the flash-lag effect. This can be observed when an airplane with a flashing tail light passes overhead and the light and the tail do not seem to line up; it can be studied in the lab with a visual stimulus illustrated in Figure 2. In the flash-lag effect, a flash and a moving object at the same location appear to be offset. One leading explanation – which makes intuitive sense, and for which there is some evidence from brain recordings – is that the brain predicts where the moving spot is going to be a short time later. However, perceptual experiments have shown that this cannot be the explanation for the flash-lag effect, because the perception attributed to the time of the flash depends on events that occur in the eighty milliseconds after the flash, not those that occur before the flash (which would be used to make a prediction).[13] This explanation for the flash-lag effect means that the brain is postdictive rather than predictive; that is, the brain is constantly revising history to make the conscious present consistent with the future. This is one example of how our brains generate plausible interpretations based on noisy and incomplete data, something that magicians have exploited for sleight-of-hand effects.[14]

Brain imaging gives us a global picture of brain activity when we perceive something compared to when we do not. Using experimental evidence, researchers have developed the particularly appealing hypothesis that we only become consciously aware of something when the level of brain activity in the front of the cortex, which is important for planning and making decisions, reaches a threshold level and ignites feedback pathways.[15] Although these observations are intriguing, they are not compelling, since they do not establish causality, only a correlation. If an NCC is responsible for a conscious state, it should be possible to change the NCC and, in so doing, change consciousness. New techniques such as optogenetics[16] have recently become available to selectively manipulate the activity of neurons, which allows the causality of the NCCs to be tested. This may be difficult to do if perceptual states correspond to highly distributed patterns of activity, but in principle this approach could reveal how perceptions and other features of consciousness are formed.

Another compelling illusion is change blindness, which can be demonstrated by altering a large object in an image, such as a parrot in a tree, during a saccade (a fast eye movement that occurs when the eye jumps from one fixation point to another). Unless a subject is paying attention to the object just before the saccade, the change will not be noticed.[17] Based on evidence from psychophysics,[18] physiology, and anatomy, philosopher Patricia Churchland, neuropsychologist V. S. Ramachandran, and I came to the conclusion in our essay "A Critique of Pure Vision" that the brain represents only what is needed at any moment to carry out the task at hand.[19] This stands in contrast to the goal of researchers in computer vision, which is to create a complete internal model of the world from an image, a goal that has proven difficult to achieve. However, a complete and accurate model may not be necessary for most practical purposes, and might not even be possible given the low sampling rate of current movie cameras.[20] The apparent modularity of vision (its relative separateness from other sensory processing streams) is also an illusion. The visual system integrates information from these other streams, in-



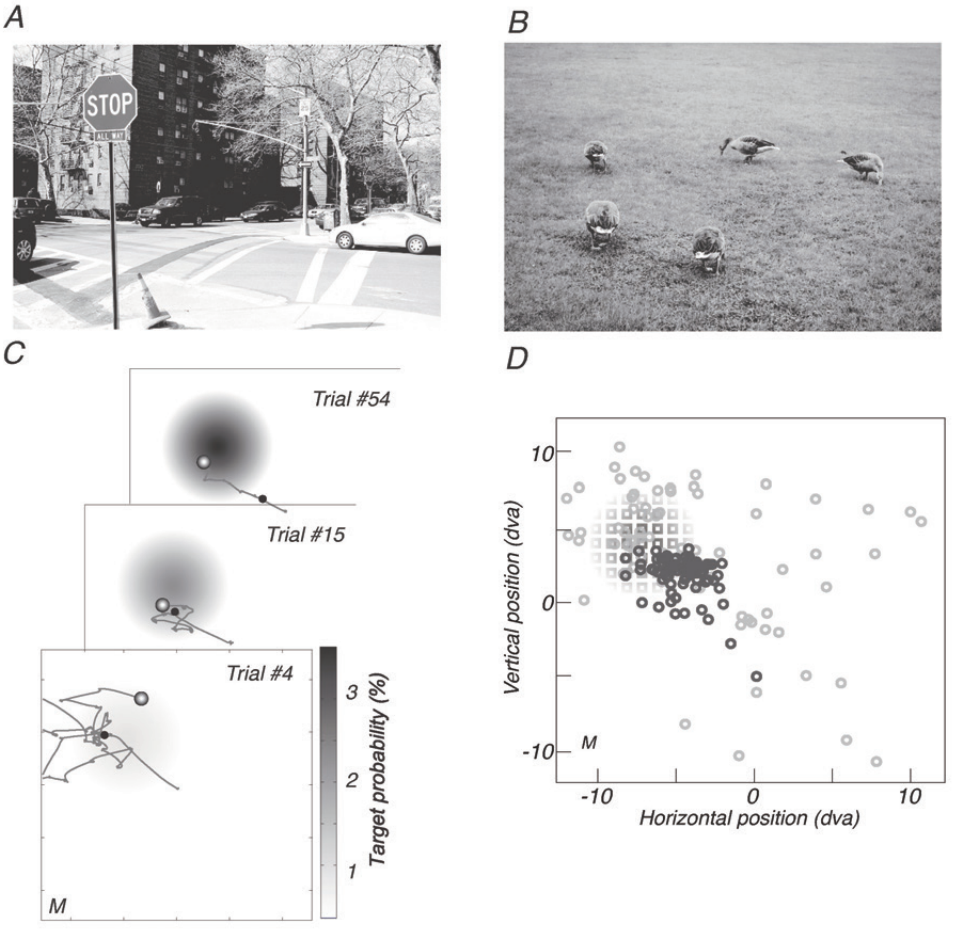
An object moves from left to right (top left). As it passes the center a light briefly flashes below it (top right). What subjects report is shown above: the object appears to be displaced to the right at the time of the flash. Source: http://hpcl.kde.yamaguchi-u.ac.jp/flashlag.html.

cluding signals from the reward system indicating the value of the scene; and the motor system actively seeks information by repositioning sensors, such as moving eyes and, in some species, moving ears.[21]

Visual search is a task that depends on both "bottom-up" sensory processing and attentional processes driven by "top-down" expectation (see Figure 3A). These two processes are intermingled in the brain and difficult to disentangle, but recently a novel search task was developed to tease them apart.[22] Participants were seated in front of a blank screen and told that their task was to explore the screen with their eyes to find a hidden target location that would sound a reward tone when their gaze fixated on it. The hidden target position varied from trial to trial and was drawn from a Gaussian distribution – a bell-shaped curve characterized by the position of its peak and width – that was not known to the participant but remained constant during a session (see Figure 3D).

At the start of a session, participants had no prior knowledge to inform their search. Once a fixation was rewarded, participants could use that feedback to assist on the next trial. As the session proceeded, participants improved their success rates by developing an expectation for the distribution of hidden targets and using it to

Figure 3
Visible and Hidden Search Tasks



(A) An experienced pedestrian has prior knowledge of where to look for signs, cars, and sidewalks in this street scene. (B) Ducks foraging in a large expanse of grass. (C) A representation of the screen is superimposed with the hidden target distribution that is learned over the session as well as sample eye traces from three trials for participant M. The first fixation of each trial is marked with a black dot. The final and rewarded fixation is marked by a shaded grayscale dot. (D) The region of the screen sampled with fixation shrinks from the entire screen on early trials (light gray circles; first 5 trials) to a region that approximates the size and position of the Gaussian-integer distributed target locations (squares, darkness proportional to the probability as given in A) on later trials (circles; from trials 32–39). Source: Leanne Chukoskie, Joseph Snider, Michael C. Mozer, Richard J. Krauzlis, and Terrence J. Sejnowski, "Learning Where to Look for a Hidden Target," *Proceedings of the National Academy of Sciences* 110 (2013): 10438–10445.

guide future searches. After approximately a dozen trials, the participants' visual fixations narrowed to the region with high target probability. A characterization of this effect for all participants is shown in Figure 3D. The search spread was initially broad and narrowed as the session progressed. Surprisingly, many of the subjects were not able to articulate their search strategy, despite the fact that after a few

trials their first saccade was invariably to the center of the invisible target distribution.[23]

The brain areas that are involved in this search task include the visual cortex and the superior colliculus, which controls the topographic map of the visual field and directs saccades to visual targets, working closely with other parts of the oculomotor system. Learning also involves the basal ganglia, an ancient part of the vertebrate brain that learns sequences of actions through reinforcement learning.[24] The difference between the expected and received reward is signaled by a transient increase in the firing rate of dopamine neurons in the midbrain, which regulates synaptic plasticity and influences how decisions and plans are made at an unconscious level.[25]

The structure of DNA was discovered in 1953 and the human genome was sequenced fifty years later. I once asked Francis Crick if he ever thought in those early years that the human genome would be sequenced in his lifetime. He said it never occurred to him that it would ever be possible. Fifty years from now, how far will we be on the problem of consciousness? By then we may have machines that interact with us in much the same way that we interact with each other, through speech, gestures, and facial expressions. However, it may be easier to create consciousness than to fully understand it. I suspect that we can make progress faster by first understanding unconscious processing: all the things that we take for granted when we see, hear, and move. We have already made progress on understanding motivational systems, which strongly influence our decisions; and attentional systems, which help guide our search for information from the world. With a deeper understanding of the brain mechanisms that govern perception, decision-making, and planning, the problem of consciousness could disappear like the Cheshire cat, leaving only a broad grin.[26]

ENDNOTES

1 Francis H.C. Crick, *What Mad Pursuit: A Personal View of Scientific Discovery* (New York: Basic Books, 1988).

2 There is no single accepted scientific definition of consciousness. However, it includes the state of being awake and aware of one's surroundings, the awareness or perception of something, and the mind's awareness of itself and the world.

3 Francis Crick, private communication to Terrence J. Sejnowski, 1998.

4 Christine Aicardi, "Of the Helmholtz Club, South-Californian Seedbed for Visual and Cognitive Neuroscience, and Its Patron Francis Crick," *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 45 (2014): 1–11.

5 Francis Crick and Christof Koch, "The Problem of Consciousness," *Scientific American* 267 (3) (1992): 10–17; Francis Crick and Christof Koch, "Are We Aware of Neural Activity in Primary Visual Cortex?" *Nature* 375 (1995): 121–123; Francis Crick and Christof Koch, "Constraints on Cortical and Thalamic Projections: The No-Strong-Loops Hypothesis," *Nature* 391 (1998): 245–250; Francis Crick and Christof Koch, "A Framework for Consciousness," *Nature Neuroscience* 6 (2003): 119–126; and Francis Crick, Christof Koch, Gabriel Kreiman, and Itzhak Fried, "Consciousness and Neurosurgery," *Neurosurgery* 55 (2) (2004): 273–281.

6 The cerebral cortex is the outer layer of the mammalian brain. It is highly convoluted in humans and is involved in memory, attention, perceptual awareness, thought, language, and consciousness.

[7] Rodrigo Quian Quiroga, Itzhak Fried, and Christof Koch, "Brain Cells for Grandmother," *Scientific American* 308 (2) (2013): 30 – 35.

[8] Karl Deisseroth and Mark J. Schnitzer, "Engineering Approaches to Illuminating Brain Structure and Dynamics," *Neuron* 80 (2013): 568 – 577.

[9] Valerio Mante, David Sussillo, Krishna V. Shenoy, and William T. Newsome, "Context-Dependent Computation by Recurrent Dynamics in Prefrontal Cortex," *Nature* 503 (2013): 78 – 84.

[10] BRAIN Initiative (Brain Research through Advancing Innovative Neurotechnologies), http://www.nih.gov/science/brain/2025/index.htm.

[11] Geoffrey E. Hinton, "How Neural Networks Learn from Experience," *Scientific American* 267 (1992): 144 – 151.

[12] David A. Bulkin and Jennifer M. Groh, "Seeing Sounds: Visual and Auditory Interactions in the Brain," *Current Opinion in Neurobiology* 16 (2006): 415 – 419.

[13] David M. Eagleman and Terrence J. Sejnowski, "Motion Integration and Postdiction in Visual Awareness," *Science* 287 (2000): 2036 – 2038.

[14] Stephen L. Macknik, Susana Martinez-Conde, and Sandra Blakeslee, *Sleights of Mind: What the Neuroscience of Magic Reveals About Our Everyday Deceptions* (New York: Henry Holt, 2010).

[15] Stanislas Dehaene and Jean-Pierre Changeux, "Experimental and Theoretical Approaches to Conscious Processing," *Neuron* 70 (2011): 200 – 227.

[16] BRAIN Initiative, http://www.nih.gov/science/brain/2025/index.htm.

[17] John A. Grimes, "On the Failure to Detect Changes in Scenes across Saccades," in *Perception* (Vancouver Studies in Cognitive Science, Vol. 5), ed. Kathleen Akins (Oxford: Oxford University Press, 1996), 89 – 110.

[18] Psychophysics is an area of psychology that deals with relationships between physical stimuli and mental phenomena.

[19] Patricia S. Churchland, V. S. Ramachandran, and Terrence J. Sejnowski, "A Critique of Pure Vision," in *Large-Scale Neuronal Theories of the Brain*, ed. Christof Koch and Joel D. Davis (Cambridge, Mass.: MIT Press, 1994), 23 – 60.

[20] Terrence J. Sejnowski and Tobi Delbruck, "The Language of the Brain," *Scientific American* 307 (2012): 54 – 59.

[21] Churchland, Ramachandran, and Sejnowski, "A Critique of Pure Vision."

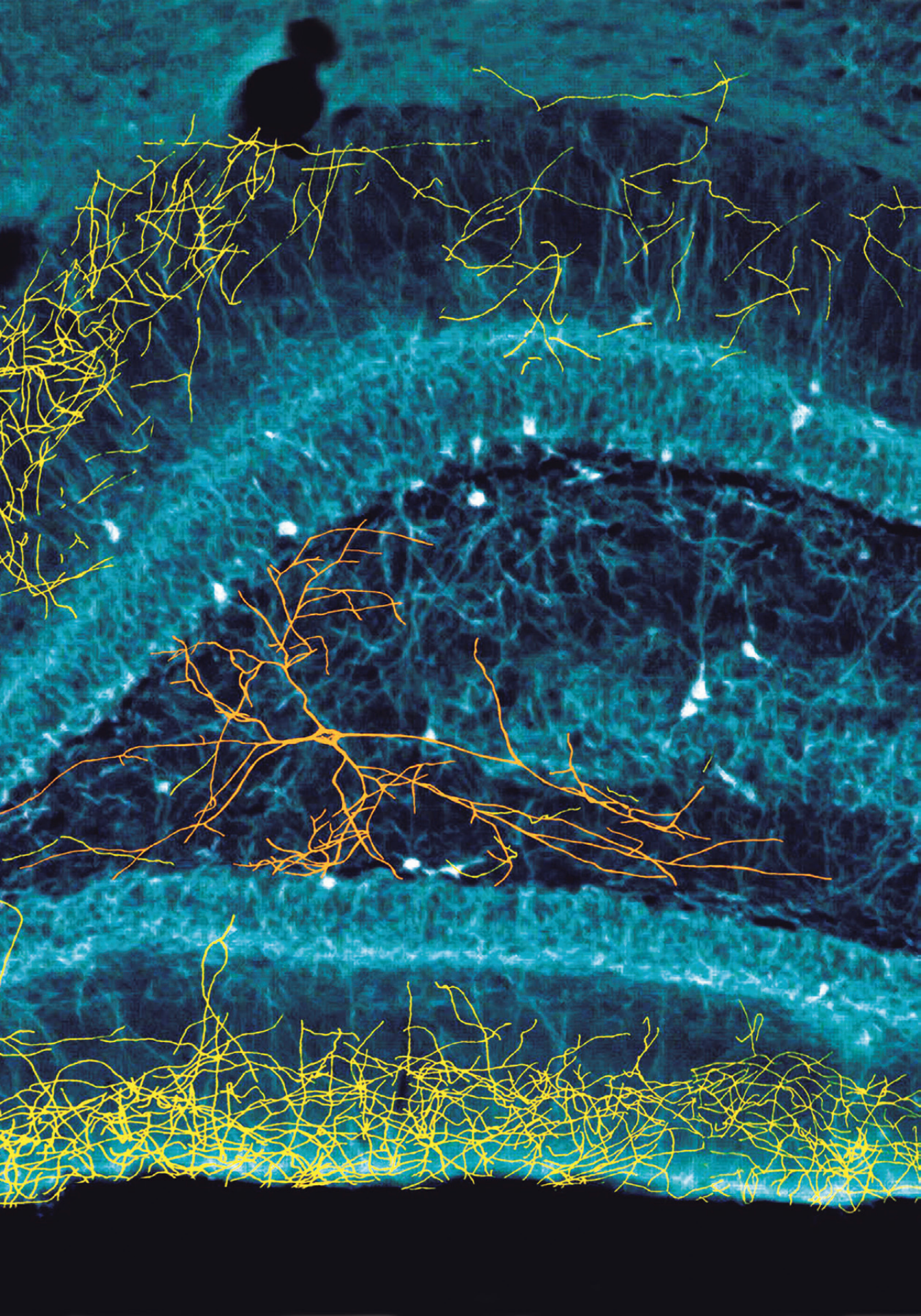
[22] Leanne Chukoskie, Joseph Snider, Michael C. Mozer, Richard J. Krauzlis, and Terrence J. Sejnowski, "Learning Where to Look for a Hidden Target," *Proceedings of the National Academy of Sciences* 110 (2013): 10438 – 10445.

[23] Ibid.

[24] Terrence J. Sejnowski, Howard Poizner, Gary Lynch, Sergei Gepshtein, and Ralph J. Greenspan, "Prospective Optimization," *Proceedings of the Institute of Electrical and Electronic Engineering* 102 (2014): 799 – 811.

[25] P. Read Montague, Peter Dayan, and Terrence J. Sejnowski, "A Framework for Mesencephalic Dopamine Systems Based on Predictive Hebbian Learning," *The Journal of Neuroscience* 16 (1996): 1936 – 1947; Wolfram Schultz, Peter Dayan, and P. Read Montague, "A Neural Substrate of Prediction and Reward," *Science* 275 (1997): 1593 – 1599; and Terrence J. Sejnowski, "Learning Optimal Strategies in Complex Environments," *Proceedings of the National Academy of Sciences* 107 (2010): 20151 – 20152.

[26] Lewis Carroll, *Alice's Adventures in Wonderland* (London: Macmillan and Co., 1865).



coming up in Dædalus:

On an Aging Society	John W. Rowe, Jay Olshansky, Julie Zissimopoulos, Dana Goldman, Robert Hummer, Mark Hayworth, Lisa Berkman, Axel Börsch-Supan, Dawn Carr, Linda Fried, Frank Furstenberg, Caroline Hartnett, Martin Kohli, Mauricio Avendano, John Rother, David Bloom, David Canning, and others
On Water	Christopher Field & Anna Michalak, Michael Witzel, Charles Vörösmarty, Michel Meybeck & Christopher L. Pastore, Terry L. Anderson, John Briscoe, Richard G. Luthy & David L. Sedlak, Stephen R. Carpenter & Adena R. Rissman, Jerald Schnoor, Katherine Jacobs, and others
Food, Health & the Environment	G. David Tilman, Walter C. Willett, Meir J. Stampfer & Jaquelyn L. Jahn, Nathaniel D. Mueller & Seth Binder, Andrew Balmford, Rhys Green & Ben Phalan, G. Philip Robertson, Brian G. Henning, and others
The Internet	David Clark, Yochai Benkler, Peter Kirstein, Deborah Estrin & Ari Juels, Archon Fung, Susan Landau, John Palfrey, and others

plus What's New About the Old?; Political Leadership; New Dilemmas in Ethics, Technology & War &c

AMERICAN ACADEMY OF ARTS & SCIENCES
Cherishing Knowledge · Shaping the Future

U.S. $14; www.amacad.org