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-

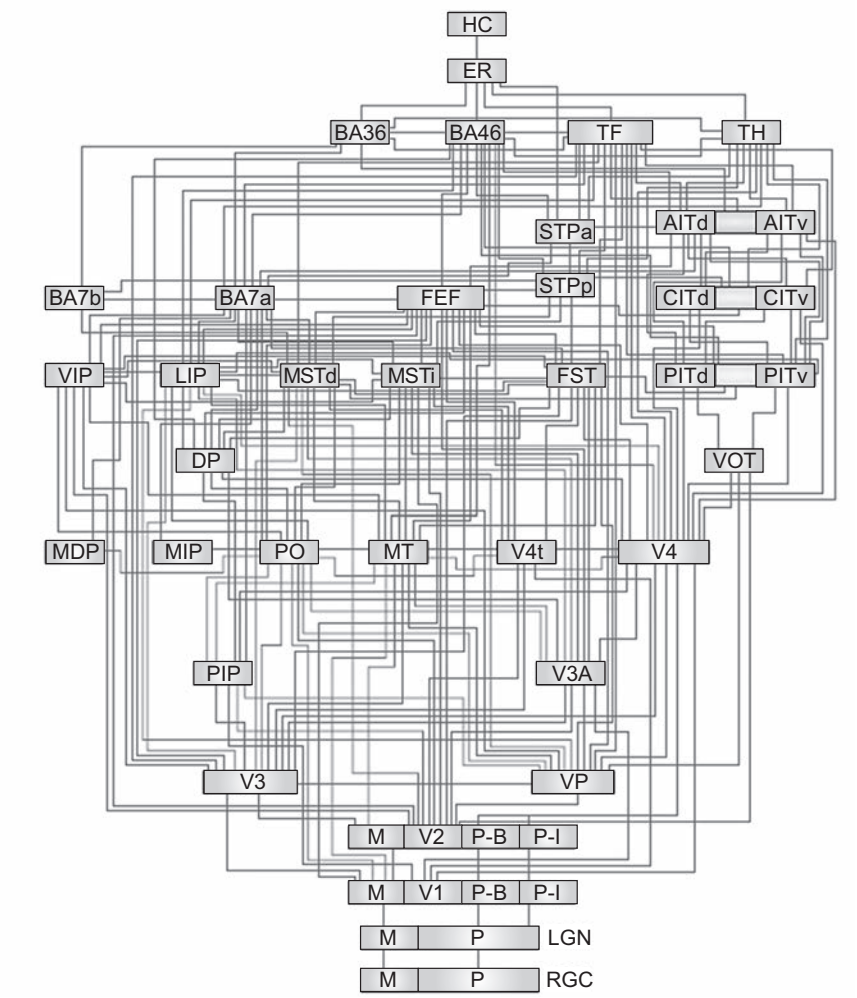
doi:10.1162/DAED_a_00321

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.

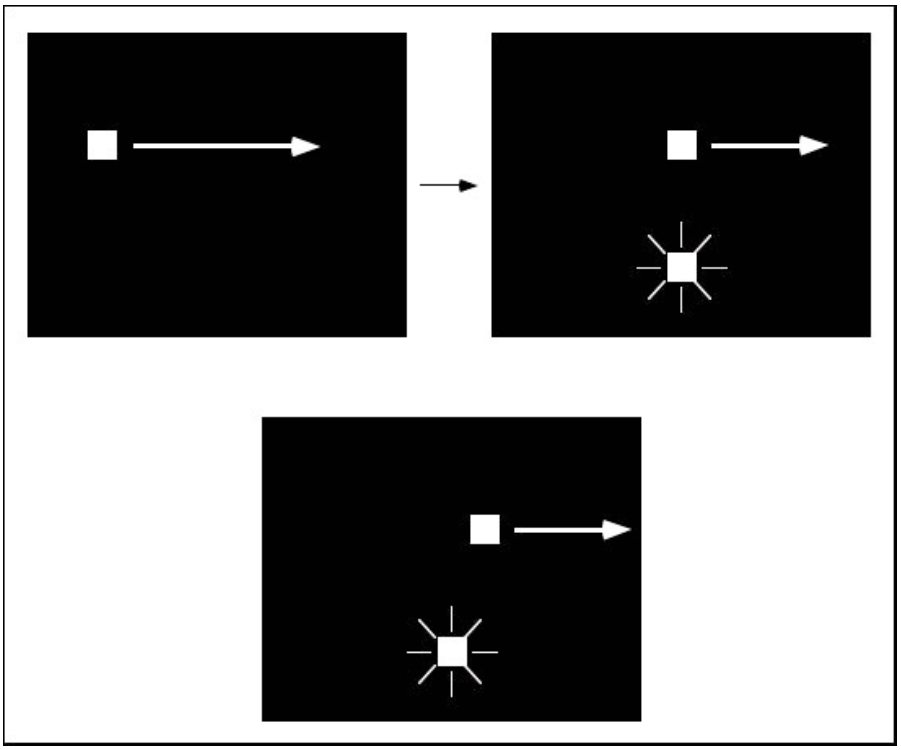
The question of simultaneity becomes even more vexing when we make cross-modal comparisons. As you are watching someone chop down a tree, you simultaneously see and hear the ax hit the tree, even though the speed of sound is much less than that of light. Moreover, the illusion of simultaneity is maintained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



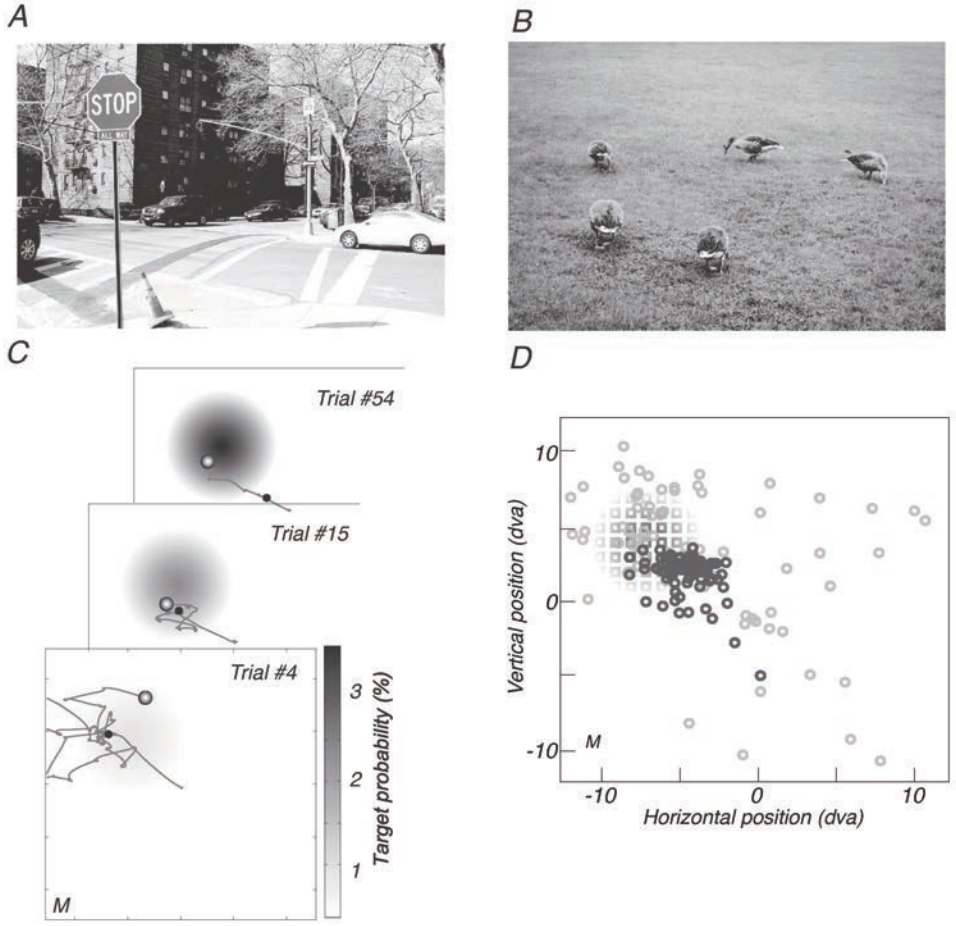
An object moves from left to right (top left). As it passes the center a light briefly flashes below it (top right). What subjects report is shown above: the object appears to be displaced to the right at the time of the flash. Source: http://hpcl.kde.yamaguchi-u.ac.jp/flashlag.html.

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).