THE GREAT COURSES®

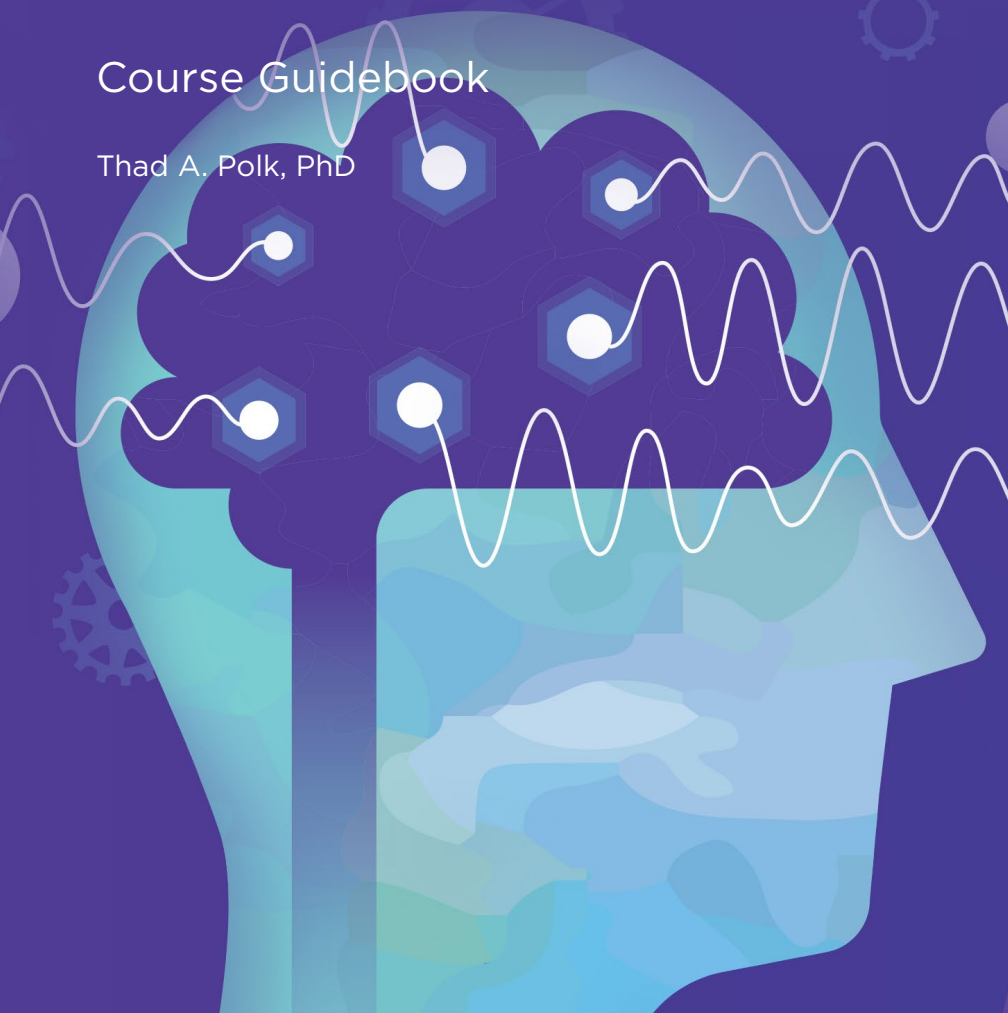
Topic
Science

Subtopic
Neuroscience
& Psychology

Introduction to Cognitive Science

Course Guidebook

Thad A. Polk, PhD



Printed in the United States of America

4840 Westfields Boulevard, Suite 400
Chantilly, VA 20151-2299
USA
1-800-832-2412
www.thegreatcourses.com

Thad A. Polk is the Samuel D. Epstein Collegiate Professor of Psychology and an Arthur F. Thurnau Professor at the University of Michigan. He earned an interdisciplinary PhD in Computer Science and Psychology from Carnegie Mellon University and received postdoctoral training in Cognitive Neuroscience at the University of Pennsylvania. His teaching has been recognized by numerous awards, and he was named to The Princeton Review's list of the Best 300 Professors in the United States.

Table of Contents

1

Opening the Black Box of the Mind

If you want to crack the mystery of human experience—of what it means to be human—the human mind is a good place to start. This course takes you on a deep dive into the fascinating field of cognitive science, which has begun to shed light on the powerful and mysterious cognitive processes that humans use every day. Topics include language, memory, decision-making, emotion, consciousness, and many other amazing cognitive processes. You'll also learn about powerful methods from artificial intelligence that are used for studying how the mind works. This lecture explores the origins of the field of cognitive science and what differentiates it from other areas of scientific study.

The Black Box Problem

As its name suggests, the field of cognitive science is the scientific study of cognition, or thought, in all its forms. Put another way, it's the scientific study of how minds work. And cognitive scientists typically aren't content with vague descriptions of mental processes. They want to know how the machinery of the brain operates. Ideally, they'd like to be able to build a simulation that executes mental processes in the same way a real mind does.

But figuring out the details of how minds work is one of the greatest challenges in all of science because there's just so much about the mind that scientists can't directly observe. This dilemma is sometimes referred to as the black box problem.

Scientists can observe the sensory inputs that are available to a mind. For example, they can observe visual inputs, such as words on a page, or auditory inputs, such as spoken language. They can also observe the outputs that a mind produces, like spoken language or motor movements. But they can't directly observe the mental processes that happen in between, such as the processes that perceive auditory input and transform them into representations of words. Nor can they directly observe the mechanisms that the mind uses to decide on and carry out specific motor actions. Science can only measure and observe physical phenomena. But cognitive processes aren't physical—they're mental.

It's true that cognitive processes are implemented by physical processes. For example, the neural networks in your brain are physical. They're made up of connections between physical neurons that carry physical electrical and chemical signals. And scientists can measure this neural activity. But the neural activity associated with a specific thought or feeling is not the same thing as the thought or feeling itself. For example, fear might be associated with activity in specific neural circuits, but fear itself is a subjective mental

To cognitive scientists, it's as if the mind is hidden inside a black box. They can see what goes into the box and what comes out of the box but not what's going on inside.

state, not the objective pattern of neural activity that gives rise to it. That makes cognitive science different from any other science, such as biology and chemistry, where physical entities can be directly observed, at least in principle.

Introspectionism and Behaviorism

Historically, one of the first approaches to studying the mental constructs involved in cognition was a technique called introspectionism. Like its root word, *introspection*, introspectionism has to do with looking inside the self. The idea was to examine one's own mind and identify, analyze, and document the most primitive atomic sensations and thoughts to gain insight into the structure and operation of the mind.

But as a scientific method, introspectionism suffered from a very serious problem, namely, that the results and conclusions were difficult to replicate and impossible to verify. The results appeal to private events to which no one but the introspectionist has access. And of course, replication and verification are at the very heart of reproducible science.

There's a lot that goes on in the mind that one doesn't have access to. Most of the cognitive processes involved in recognizing a face, retrieving a relevant fact, speaking a word, and countless other cognitive tasks happen automatically and unconsciously. And unfortunately, introspection can't provide much insight into any of those processes.

Partly in response to these concerns, most scientists who were interested in the mind began adopting a fundamentally different approach in the early 1900s. That approach came to be known as behaviorism.

Behaviorists believed that behavior could be understood and predicted by observing external, objective phenomena. They examined relationships between directly observable stimuli and responses.

Behaviorists wanted to develop a rigorous scientific approach to studying the mind, so they left behind speculation about subjective mental processes. Instead, they focused on studying observable behavior. Behaviorists often taught animals, such as rats, to associate food rewards with specific sensory stimuli or behaviors. Once the animal had been trained, the researchers would examine how it responded and changed its behavior based on the association between a stimulus and a reward. For example, Ivan Pavlov famously investigated how pairing the sound of a bell with a reward like food led dogs to associate the bell with the food—sometimes referred to as Pavlovian or classical conditioning.

B. F. Skinner found that behaviors that were reinforced by food were more likely to be repeated in the future. And this kind of so-called operant conditioning was so effective that Skinner went on to hypothesize that most aspects of human behavior were acquired in a similar way.

The behaviorist approach led to significant progress in the understanding of learning and memory and the role of reinforcement. Behaviorism dominated scientific investigations of the mind in the first half of the 20th century.

The Cognitive Revolution

In the 1950s, a number of events conspired to lead scientists back to that black box of cognition. It was beginning to seem possible to develop detailed models of the mental processes and structures underneath human cognition. The change in outlook was so significant that scientists now often refer to this period as the cognitive revolution.

One of the most important developments was a criticism of behaviorism that was published by the influential linguist Noam Chomsky. Chomsky argued that behaviorist ideas like operant conditioning couldn't explain the human ability to acquire complex natural languages such as English, Chinese, and Hindi.

For example, people regularly produce sentences that they've never produced before, and so those sentences could not have been reinforced in the past. That's a problem for behaviorist theories that assume the most important learning mechanism is the reinforcement of past behavior.

The invention of the digital computer was perhaps the single most important development in the cognitive revolution and the rise of cognitive science.

Chomsky persuasively argued that if scientists ever wanted to understand language, they were going to have to go beyond directly observable stimuli, responses, and reinforcements and begin to look inside the black box of cognition. But is there any way scientists can develop theories about unobservable mental processes while maintaining scientific rigor?

Enter the digital computer. Although the connection might not seem obvious at first, the digital computer provided a new metaphor for thought that was flexible and powerful, as well as rigorous. When they were first introduced, digital computers were primarily thought of as powerful calculators. These machines could crunch numbers quickly and accurately using only 0s and 1s.

But scientists quickly realized that computers could be used as more general, programmable information processors. They saw that 0s and 1s didn't exclusively have to represent numbers; they could also be used to represent arbitrary symbols. And those symbols could be used to represent virtually any concept or idea. Such a realization changed not only the way scientists viewed computers but also the way they thought about the human mind: Maybe a mind is something like an information processing computer.

Symbols are arbitrary representations that are assigned or associated with a specific meaning. For example, the word *cat* in English is a symbol that is associated with a four-legged domesticated animal that purrs and is commonly kept as a pet. There's nothing special about the word itself that reflects its meaning. Other languages adopt different conventions about which symbols refer to that animal.

Many cognitive scientists would argue that there's nothing magical about minds and that scientists will eventually figure out exactly how they work at a computational level. Other cognitive scientists believe that some aspects of mental experience can't be reduced to computational mechanisms.

In particular, some cognitive scientists have argued that subjective, conscious experiences can't be explained by physical mechanisms. That said, virtually every cognitive scientist would agree that the computational view of the mind has had a profound and ongoing influence on the field.

Computational Theories

Cognitive scientists regularly develop computational theories of memory, language, attention, vision, decision-making, and other aspects of cognition. In fact, they often implement real computational simulations of those processes that can be run on a laptop computer. These models can then be run under many different conditions to generate explicit predictions that can be compared to real empirical data. And that provides a way to figure out how the black box of the mind actually works without falling prey to the problems associated with introspectionism.

Computationally explicit theories make verifiable and testable predictions. For example, if one cognitive scientist implements a computational simulation of memory or language, others can also run that simulation and verify the predictions. They don't have to take someone else's word for it. Furthermore, those predictions can then be compared to data from real experiments about memory or language. And so it becomes possible to test whether a theory's predictions are correct. Perhaps even more importantly, it makes it possible to demonstrate that a theory is wrong and needs to be modified. Consequently, researchers can engage in the self-correcting process of science, in which theories are proposed, compared against empirical data, and then revised or replaced in light of that data.

Of course, developing plausible computationally explicit theories of how the mind works is no easy task! And neither is designing experiments to test those theories. So cognitive scientists draw inspiration from quite a few different fields, including psychology, artificial intelligence, neuroscience, philosophy, linguistics, and anthropology. Cognitive science is an extremely interdisciplinary field of study, and this course explores studies and ideas from almost all of these areas.

Reading

Bermúdez, J. L. *Cognitive Science: An Introduction to the Science of the Mind*. 3rd ed. Cambridge: Cambridge University Press, 2020.

Miller, G. A. "The Cognitive Revolution: A Historical Perspective." *Trends in Cognitive Sciences* 7, no. 3 (2003): 141–144. https://doi.org/10.1016/S1364-6613(03)00029-9.

Pinker, S. *How the Mind Works*. New York: Norton, 1997.

2

A Hands-On Guide to Brain Anatomy

Understanding the basics of brain anatomy will be very helpful as you go through the rest of course. This lecture provides an overview of the structures of the brain by dissecting some real brains, examining the parts, and talking about what those parts do and how they contribute to cognition. One brain is from a generous adult human who wanted to contribute to science. The other brain is from a sheep, and although the sheep's brain is smaller than the human brain, it shares many of the same structures.

Directional and Orientational Terminology

If you want to dissect a real brain at home, you can buy a sheep brain online for about $15. You may even want to get two or three so that you can cut each brain in a different direction. If you do your own dissection, though, here's a word of warning: Sheep brains are typically preserved in formalin, which smells bad and can irritate your eyes, nose, and throat. So you should wear rubber gloves while you do the dissection and avoid touching your face until you've removed the gloves and washed your hands. You can also find lots of labeled pictures of sheep-brain dissections on the internet.

To navigate the brain, you first need to understand some crucial directional terminology. *Anterior* means "toward the front," and *posterior* means "toward the back." You'll also sometimes hear the terms *rostral*, which literally means "toward the beak" or "toward the nose," and *caudal*, which means "toward the tail." So in the brain, *rostral* and *anterior* mean the same thing: "toward the front." Likewise, *caudal* and *posterior* mean the same thing: "toward the back."

Superior means "toward the top," and *inferior* means "toward the bottom." Superior parts are sometimes referred to as being dorsal, which literally means "toward the back." For example, you've probably heard of the dorsal fin of a shark or dolphin, which is a fin on the animal's back—the one you might see sticking out of the water while it swims around.

Conversely, the term *ventral* is sometimes used to refer to the inferior part of the brain. Whereas *dorsal* means "back," *ventral* means "belly." So *ventral* means "toward the bottom," just like *inferior*.

Lateral means "toward the side," and *medial* means "toward the middle." If looking at a brain face on, moving laterally means moving toward the left or right side, and moving medially means moving from either side toward the middle.

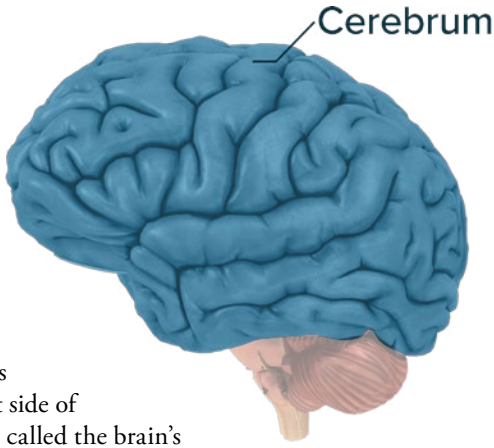
It's also helpful to distinguish different orientations when you're looking at a brain. For example, if you're looking at the brain from the side, that's called a sagittal view. If you're looking at a midsagittal section, that means you're looking at the middle of the brain from a side view.

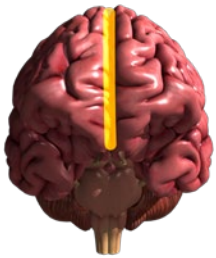
If you're looking at the brain from the front or the back, that's called a coronal view. Finally, if you're looking at the brain from the top or bottom, that's called an axial view. Axial sections are also sometimes called horizontal sections because they're horizontal or parallel to the ground.

The Cerebrum

At the largest scale, the brain can be divided into three major parts: the cerebrum, the cerebellum, and the brain stem.

The cerebrum is the biggest part of the brain and is divided into two cerebral hemispheres by a deep and long valley called the longitudinal fissure. In general, the left hemisphere controls the right side of the body and processes information coming from the right side of space. Conversely, the right hemisphere controls the left side of the body and processes information coming from the left side of space. This makeup is sometimes called the brain's contralateral organization.




Longitudinal fissure

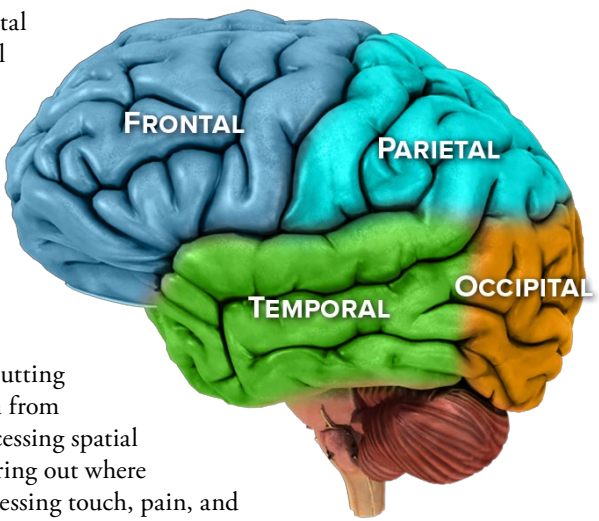
The cerebral hemispheres have a bunch of hills that are separated by valleys. Each hill, or ridge, is called a gyrus, and each valley, or groove, is called a sulcus. Inside the brain, the outer layer of the cerebral hemispheres is different from the interior of the cerebral hemispheres. The outer layer is called the cerebral cortex and is composed of gray matter rather than white matter. The cerebral cortex is kind of like the bark around the trunk of the tree and is where most of the neural information processing takes place.

In most parts of the human brain, the cerebral cortex is composed of six layers of tissue and is like a sheet around the outside of the brain. So imagine laying out a flat sheet of cells on a table and then trying to cram that sheet into a spherical skull. You'd have to crunch it up like you would a flat piece of paper to make it into a ball. And when you do that, the result is a bunch of ridges and valleys, which is exactly what's going on with the ridges and valleys in the brain.

The cerebral cortex is where a lot of the action is in terms of higher-order thought. The cortex is where language gets processed, where sights and sounds get recognized, and where planning and problem-solving take place. The cerebral cortex is also where your personality resides.

The cerebral hemispheres are typically divided into four lobes: frontal, parietal, occipital, and temporal. The frontal lobes are the most anterior part of the cerebral hemispheres. They play a role in many behavioral functions, but three of the most important ones are language production, voluntary motor control, and executive function, which is involved in deciding what to do or think about next.

Immediately behind, or posterior to, the frontal lobes, are the parietal lobes, which are toward the top and the back of the brain. They, too, play a role in many behavioral functions, but three of the most important ones are putting together information from different senses, processing spatial information (or figuring out where things are), and processing touch, pain, and temperature.

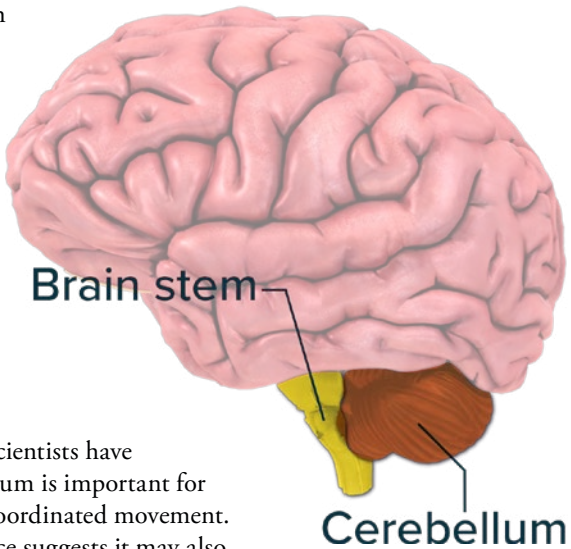


At the very back of the brain is the occipital lobe, which is the most posterior part of the cortex. It is critically important in processing light, color, and motion and is therefore crucial for vision.

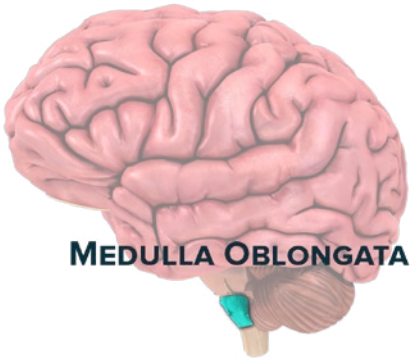
Finally, on the two sides of the brain, inferior to the parietal and frontal lobes and anterior to the occipital lobe, are the temporal lobes. These lobes are the home of auditory cortex, where sounds get processed and recognized. They also play a critical role in language comprehension. A lot of the brain regions involved in visual recognition reside in the temporal lobe, as do many regions involved in long-term memory.

The Cerebellum and the Brain Stem

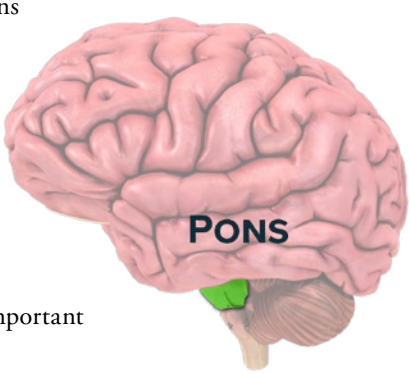
The structure at the back of the brain that looks kind of like cauliflower is the cerebellum, which is Latin for "little brain." The name presumably derives from the fact that the cerebellum makes up only about 10% of the brain's volume. Nevertheless, it contains more than 50% of all the neurons in the brain.



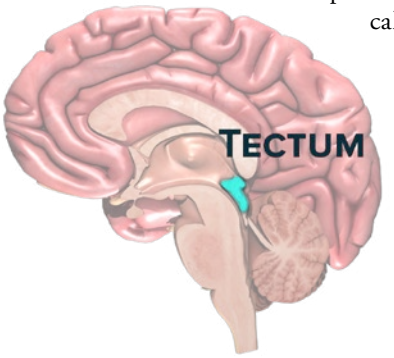
For a long time, neuroscientists have known that the cerebellum is important for balance, posture, and coordinated movement. But more recent evidence suggests it may also play an important role in some cognitive and emotional functions.



The other major part of the brain that you can see from the outside is the brain stem, which plays a crucial role in the regulation of unconscious bodily functions. Things like breathing, heart rate, and body temperature all depend on the brain stem. It consists of three major parts. The most posterior, or caudal, part is the medulla oblongata, which continues down into the spinal cord. In addition to containing centers for breathing, heart rate, and blood pressure, the medulla also contains regions critical for vomiting, coughing, sneezing, and swallowing.



The part of the brain stem that swells out a bit is called the pons, which comes from the Latin word for "bridge." In addition to containing neural pathways that send signals between the cerebrum and the cerebellum, the pons is also crucially important in sleep and dreaming.



The most anterior and superior part of the brain stem is called the midbrain. You can't really see it from the outside, but if you separate the cerebellum from the cerebral hemispheres, then you can look in between them and see the top of the midbrain, which is called the tectum. *Tectum* means "roof" in Latin, which is appropriate because it is the roof of the midbrain.

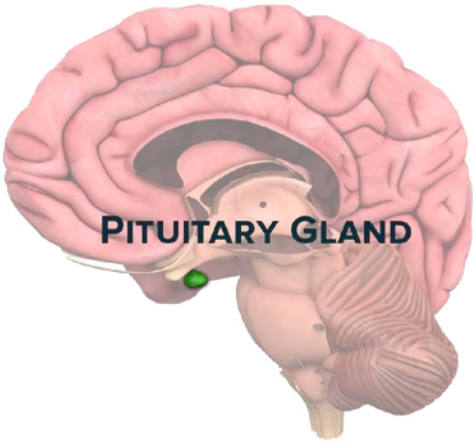
If you look carefully at the tectum, you might be able to see two pairs of bumps or mounds. The larger pair of bumps on the top are the superior colliculi, which means "upper hills" in Latin. The superior colliculi are the main control center for eye movements.

Immediately below the superior colliculi, you might be able to see a pair of smaller bumps. Those are the inferior colliculi, or "lower hills," and they're a major relay station for sound information on the way from the ears to the brain.

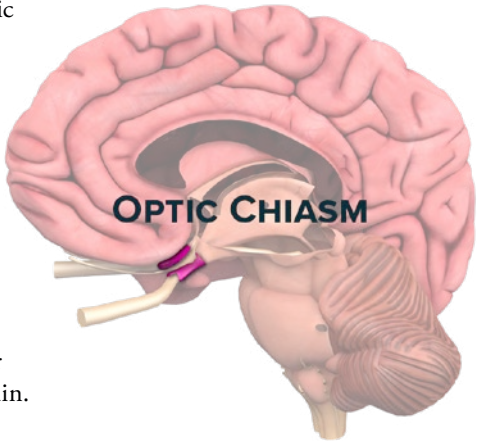
The Ventral Surface of the Brain

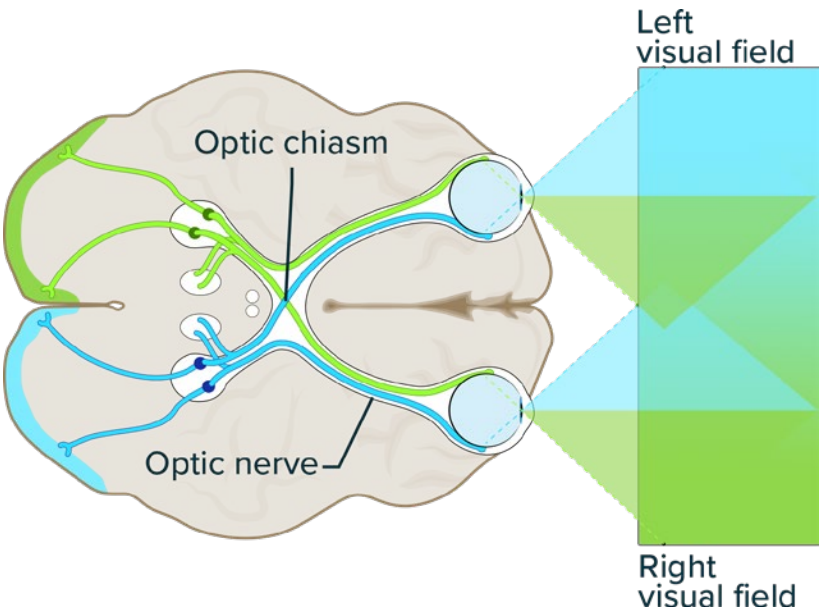
When you look at the ventral surface of the brain, one prominent feature you might see is a kind of bulb protruding from the middle of the surface. That's the pituitary gland. Unfortunately, it often gets ripped off when the brain is extracted from the skull or when the dura mater is removed. The dura mater, which literally means "tough mother," is one of the three meninges that surround and protect the brain and spinal cord.

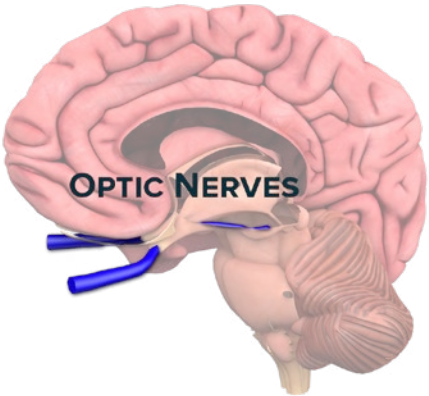
The pituitary gland secretes a number of important hormones into the bloodstream, including stress hormones that are involved in the fight-or-flight response, sex hormones that stimulate ovulation in women and sperm production in men, and growth hormones.



Just anterior to the pituitary gland is an X-shaped structure called the optic chiasm, from the Greek word for "crossing." Recall the brain's contralateral organization—the left cerebral hemisphere processes information on the right side of space, and the right hemisphere processes information on the left side of space. The optic chiasm is where information coming from the eyes crosses over to the contralateral side of the brain.

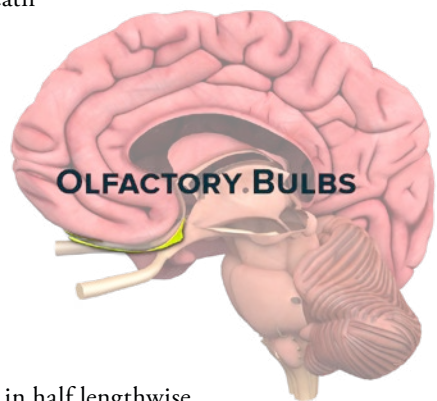






The two stalks that come into the optic chiasm are the optic nerves, which consist of bundles of fibers coming from each eye. Both eyes process information from the left and right sides of space, which means that only half of the fibers cross over. The fibers from the left eye corresponding to the left side of space cross over to the right hemisphere, and the fibers from the right eye corresponding to the right side of space cross over to the left hemisphere.

The two large stalks that are underneath the frontal lobe are the olfactory bulbs, which process smell. In humans, they sit on top of the sinuses, where they can receive smell information from the nasal cavities.



Internal Structures of the Brain

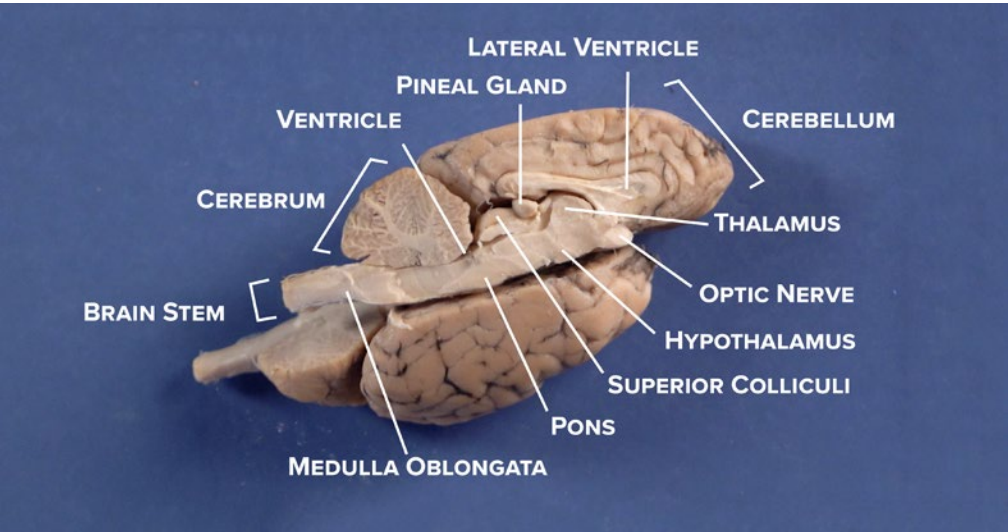
One way to dissect a brain is to cut it in half lengthwise along the longitudinal fissure that separates the left and right hemispheres, which will allow for a midsagittal view of the brain. Recall that a sagittal view is a view from the side, so a midsagittal view is a view of the middle of the brain from the side.

From this view, it's possible to see inside the three main structures discussed in this lecture: the cerebral hemispheres, the cerebellum, and the brain stem, including the medulla oblongata, the pons, and the midbrain. When looking at a complete cross-section through the midbrain, you can see such features as the superior and inferior colliculi of the tectum.

Also from this view, it's possible to see the brain's ventricular system. You've probably heard of a lumbar puncture or spinal tap, where a needle is inserted between two lumbar bones in the spine to extract cerebrospinal fluid. That fluid circulates not only around the spinal cord but also in the brain. And the fluid-filled cavities in which the cerebrospinal fluid circulates are called ventricles. Each cerebral hemisphere has its own lateral ventricle, and there's a ventricle between the brain stem and the cerebellum.

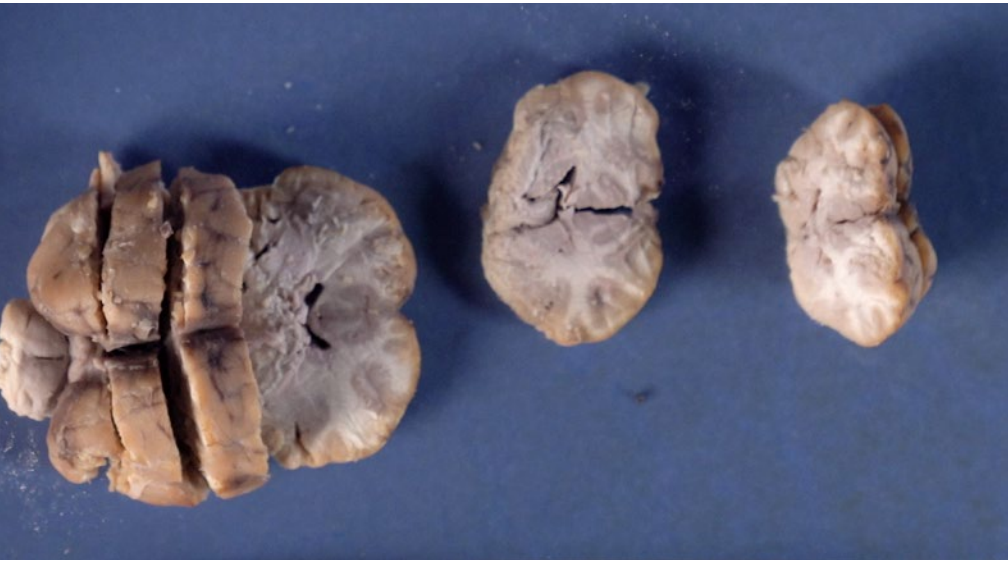
Immediately above, or superior to, the lateral ventricles is a whitish band of fibers called the corpus callosum, which is the major set of fibers connecting the left and right cerebral hemispheres. Just below the most posterior part of the corpus callosum and just anterior to the superior colliculi is the pineal gland, also called the pineal body. It produces the hormone melatonin, which helps to modulate sleep patterns.

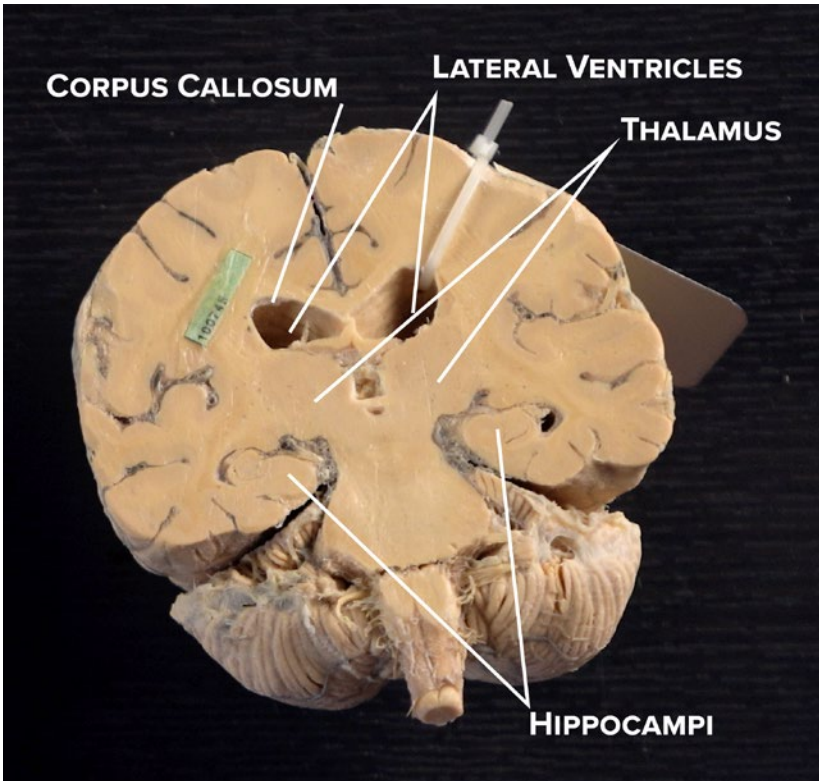
Just anterior to the pineal gland is a larger circular structure in the middle of the brain. This structure is the thalamus, which is the brain's major sensory relay station. Visual information from the eyes, auditory information from the ears, touch information from the skin, and taste information from the mouth are sent to the thalamus for processing before being relayed to sensory areas in the cerebral cortex. The thalamus also plays a critical role in regulating consciousness, sleep, and alertness. And immediately below, or inferior to, the thalamus is the hypothalamus, which literally means "under the thalamus."



Another way to examine a dissection is by looking at coronal sections, which are thinly sliced sections, almost like a loaf of bread. It's important to keep the slices in order so that you know where each slice is in relation to the rest of the brain.

If you look at a coronal slice in the frontal lobe, one of the first things to notice is the layer of gray matter around the outside of the brain. That's the cerebral cortex—the bark around the outside of the brain's trunk. And inside that is a lighter white matter that looks like tree branches. The white matter consists of communication fibers connecting different parts of the brain, especially different areas of the cerebral cortex.





In the human brain, there's one particularly prominent band of white matter fibers going from the left cerebral hemisphere to the right cerebral hemisphere. That's the corpus callosum. And just inferior to the corpus callosum are two cavities. Those are the lateral ventricles that contain cerebrospinal fluid, or CSF.

It's also possible to see the ridges and grooves of the cerebral cortex. Recall that each of the ridges is called a gyrus, and each of the grooves is called a sulcus. You can also see some gray matter immediately inferior to the lateral ventricles that is not on the outside of the brain; that is, it's not part of the cerebral cortex. This subcortical gray matter is called the basal ganglia, and, like the cerebral cortex, they also do some important information processing.

In particular, they seem to be critically important in deciding what movements to make, or what is sometimes called action selection. In fact, Parkinson's disease, which is characterized by motor problems, is associated with dysfunction in the basal ganglia.

The basal ganglia have also been implicated in motivational processing associated with wanting and liking. In fact, habit-forming drugs, like heroin and cocaine, are known to impact neural circuits between the midbrain and the basal ganglia and to produce the strong cravings associated with addiction.

If you look at a coronal section from a human brain that is slightly more posterior and that goes through the thalamus, you can still see the cerebral cortex around the outside of the brain and the branches of white matter inside that. You can also see the corpus callosum connecting the two hemispheres and the ventricles immediately inferior to that. And inferior to the ventricles, you can see the two circular-shaped sides of the thalamus, which is the brain's major relay station that sends visual, auditory, taste, and touch information to different parts of the cerebral cortex.

Also in this section are spiral-like structures in the medial part of the temporal lobes. Those are the hippocampi, one on the left and one on the right. The hippocampus plays a critical role in storing new long-term memories, and damage to it can lead to amnesia.

This course will reference a lot of the structures you learned about in this lecture, so feel free to refer back to it if you need a refresher.

Reading

Carter, R. *The Human Brain Book: An Illustrated Guide to Its Structure, Function, and Disorders*. New York: DK Publishing, 2019.

Felten, D. L., M. K. O'Banion, and M. S. Maida. *Netter's Atlas of Neuroscience*. Philadelphia: Elsevier, 2016.

Vanderah, T. W., and D. J. Gould. *Nolte's the Human Brain: An Introduction to Its Functional Neuroanatomy*. Philadelphia: Elsevier, 2021.

3

How We Acquire and Understand Language

Humans produce and comprehend millions of different sentences in more than 7,000 different languages. And they do this so effortlessly and so fast that it's easy to overlook just how amazing this ability is. But if you've ever tried to learn a second language, you may have gained an appreciation for just how complex natural languages are and how incredible it is that people can use them so fluently. This lecture is the first of three that focus on language, a cognitive process that distinguishes human beings from any other species. In this one, you'll learn about some of the key features of natural languages and how humans acquire and comprehend them.

Language Acquisition

To begin the discussion of what an absolutely amazing intellectual accomplishment learning a natural language really is, let's consider what it must be like for a baby. Babies hear a lot of sounds: traffic, dogs, thunder, the dishwasher, and many others. They also sometimes hear people speaking one or more natural languages, which they will eventually learn themselves. But before they can do that, they need to be able to distinguish language sounds from all the other sounds in their environment.

Once they can distinguish language sounds, they need to be able to parse those sounds into their component parts, which are called phonemes. Phonemes are the smallest unit of language sound. For example, the word *cat* consists of three phonemes: /k/, /ah/, and /t/. There are usually only a few dozen different phonemes in a given language, but different languages use different phonemes. And babies must learn to quickly recognize the specific phonemes used in their language.

By age 5, most children can understand at least 10,000 different words. By age 10, most children have internalized the rules for their language.

Next, they need to figure out which phonemes are grouped together in words and which phonemes are in different words. That isn't easy to do at all because normal speech consists of an extremely rapid stream of phonemes with no clear breaks between words.

Next they need to figure out what the words mean. But words are just arbitrary noises that typically don't provide any hints as to what they mean. Worse yet, words are often ambiguous. The same noise can refer to different things. Conversely, different words can also refer to the same thing, which obviously complicates the problem of figuring out what words mean.

But children also need to learn the many different rules of their language, as well as the myriad exceptions to those rules. They need to learn to form the plural form by adding *-s* to many nouns but not all nouns. They need to learn

to form the past tense by adding *-ed* to many verbs but not all verbs. They need to learn the extremely complicated rules that determine how words can be combined into phrases and phrases into sentences.

What's even more amazing is that children learn these extremely complicated natural languages without a lot of negative feedback. A number of linguists have investigated whether children learn a language faster if they are told when they make mistakes, and they found that it didn't make any difference. So, many cognitive scientists argue that human beings must be genetically hardwired to learn natural language. They don't need any kind of explicit instruction; they're just born to learn language simply by listening to the language they hear in their environment as children.

Language Comprehension

Predicting

People tend to produce four to five syllables every second and more than 100 words per minute. And the sentences that people produce are typically completely novel combinations of words and phrases that the listener needs to comprehend in real time. So how do people process the information so fast?

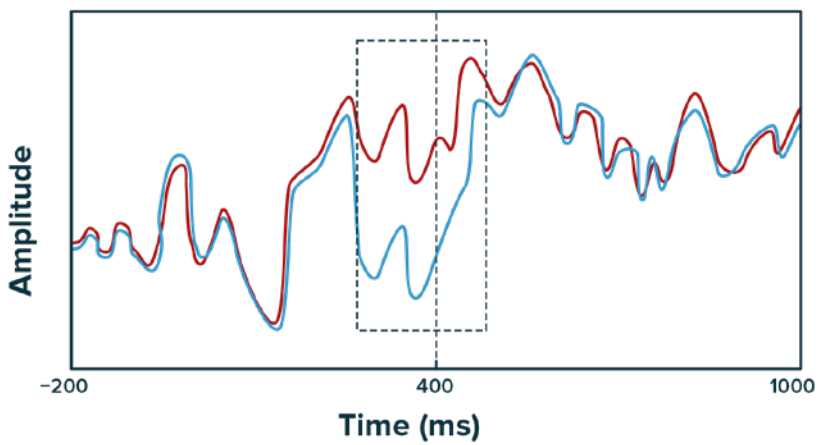
The truth is that scientists still don't entirely know. But one thing they do know is that humans are constantly making guesses about what's coming next and about the correct interpretation of the sentences they're hearing. Most of the time, they guess correctly though they're completely unaware that they were guessing. But occasionally, people guess incorrectly and when they do, they're often very aware of their mistake.

Imagine someone says to you, "The pizza was too hot to … ." Before you even hear the last word, your language circuits will already be predicting that the next word will be something like *eat* or *consume* or *carry*.

Unexpected words cause a noticeable and consistent spike in negative brain activity.

Scientists know this happens because if the sentence ends with a word that doesn't fit, you will exhibit significant brain activity—called an N400—when that unexpected word shows up. The N400 is what's called an event-related brain potential, or ERP. It was discovered by Marta Kutas and Steven Hillyard at the University of California, San Diego. They asked participants to read a bunch of sentences while they wore a cap containing electrodes that measured brain waves. Some of the sentences ended in words that you might expect, but other sentences ended in words that didn't make sense given the rest of the context.

Kutas and Hillyard found a consistent pattern of brain activity, both for sentences that ended in expected words and for sentences that ended in unexpected words. Roughly 400 milliseconds after the final word was presented, they saw a sharp spike in negative voltage in some of the electrodes—but only when the sentence ended in an unexpected word. That's the N400 ERP. The *N* stands for "negative," and the number 400 refers to the fact that this greater negativity appears around 400 milliseconds after the unexpected word is presented.



Priming

While guesswork usually works great, interpreting sentences becomes difficult when having to deal with ambiguous words, which raises another central question in the cognitive science of language: Do people access only the context-appropriate meaning of an ambiguous word, or do they access all the meanings and then narrow it down? David Swinney conducted a famous experiment to address that very question when he was at Tufts University.

Participants listened to sentences using headphones. While they listened, they also had to simultaneously perform a visual task. Letter strings were flashed on a computer string, and whenever they saw one, they had to indicate whether it was a word as quickly as possible.

For example, participants might have heard this sentence:

The man was not surprised when he found several spiders, roaches, and other bugs in the corner of his room.

Immediately after they heard the ambiguous word *bugs*, Swinney would sometimes flash a word on the screen that was related to *bugs*, like the word *ant*. When he did that, he found that people were faster to say that *ant* was a word than to say that an unrelated word, such as *sew*, was a word.

The above is an example of semantic priming. People were faster at processing *ant* when a semantically related word, like *bugs*, preceded it. *Bugs* primed the pump for *ant* and made processing *ant* easier and faster.

The same thing happened even when Swinney flashed a word that was related to one of the other meanings of *bugs*. For example, people were faster to say that *spy* was a word if they had just heard the word *bugs* in the sentence, even though *bugs* referred to insects in the sentence and not the listening-device type of bug that a spy would use.

The results suggest that people access all the different potential meanings of an ambiguous word as soon as they hear it. Afterward, they identify the specific meaning that best fits the context.

Processing Semantics and Syntax

Semantics refers to meaning. When people communicate, they need to understand the meaning of the words they hear, but they also need to know how to figure out the meaning of sentences from the meanings of the component words. Syntax refers to the grammatical rules of a language.

When people comprehend language, they have to do both syntactic and semantic processing. So do those two types of processing happen separately, or do they interact? For example, do people perform all the syntactic (grammatical) processing first and only afterward figure out what the words and sentences mean? Or are people constantly doing both syntactic and semantic processing during language comprehension?

Dan Slobin performed a famous experiment as part of his dissertation at Harvard to address this question. Children heard a sentence and were then presented with a picture. And they had to indicate as quickly as they could whether the sentence described the picture.

Slobin found that the kids were significantly faster at verifying the sentence *The cat is chasing the dog* than they were at verifying the sentence *The dog is being chased by the cat*. This makes sense because the first sentence is a syntactically simple active-voice sentence, whereas the second one is a syntactically more complex passive-voice sentence. So even though both sentences matched the picture and conveyed the same basic semantics, the kids processed the sentence with the simpler syntax more quickly.

However, in another test, the children were just as fast at verifying the passive-voice sentence *The flowers are being watered by the girl* as the active-voice sentence *The girl is watering the flowers*. So even though the first sentence was syntactically more complicated than the second one, it didn't slow the kids down in this case. But why?

One natural explanation is based on the semantics of the sentences. In the sentence *The girl is watering the flowers*, there's only one plausible agent doing the watering, namely, the girl. That is, it doesn't make semantic sense to say that the flowers are watering the girl. And because the sentence is semantically nonreversible, it's easy to figure out who is doing what to whom, even if people hear the passive-voice version.

But the original sentence—*The cat is chasing the dog*—is semantically reversible, which means the sentence *The dog is chasing the cat* makes perfect sense. And so, for the original sentence, the kids had to use the syntax of the sentence to figure out whether it matched the picture. They couldn't rely on their semantic knowledge about who typically does what to whom. And because the sentence *The dog is being chased by the cat* is more syntactically complicated, the kids took longer to match the sentence with the picture.

These results tell researchers something important about the way humans comprehend language. Syntax and semantics interact all the time during language comprehension rather than working independently.

Semantics refers to processing the meaning of words and sentences in language. Syntax refers to how words can be combined into phrases and how phrases can be combined into sentences in grammatically legal ways. People constantly use both semantics and syntax to help them comprehend the language they hear or read.

Reading

Harley, T. A. *The Psychology of Language: From Data to Theory*. London: Taylor & Francis, 2013.

Kuhl, P. K. "Baby Talk." *Scientific American* 313, no. 5 (2015): 64–69. https://doi.org/10.1038/scientificamerican1115-64.

Pinker, S. *The Language Instinct: How the Mind Creates Language*. New York: W. Morrow and Co, 1994.

Sedivy, J. *Language in Mind: An Introduction to Psycholinguistics*. 2nd ed. New York: Oxford University Press, Sinauer Associates, 2019.

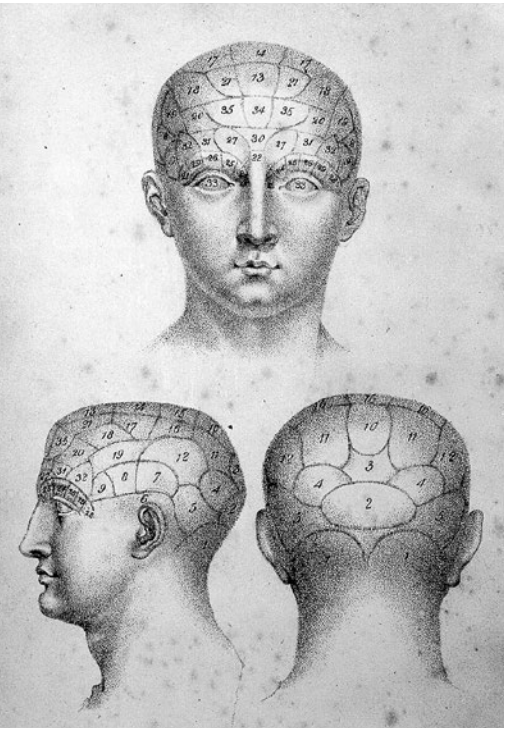
4

The Neuroscience of Language

This lecture explores how language is implemented in the human brain. Specifically, it examines the hypothesis that particular areas of the brain might be more important for language than other areas. You'll learn about important contributions from Jean-Baptiste Bouillaud, Ernest Auburtin, Paul Broca, and Carl Wernicke. You'll also learn about models that expanded on their work, including the Wernicke–Geschwind model and the dual-stream model.

Phrenology

In the early 1800s, a German physician named Franz Joseph Gall and his close collaborator Johann Gaspar Spurzheim began promoting the idea of phrenology. The basic idea was that individual differences in the shape of the skull were related to individual differences in cognitive abilities and personality traits. Phrenologists would carefully feel the skulls of their patients, looking for bumps and indentations. The size of a bump was assumed to correspond to how much the patient used that "brain organ."



Phrenology became very popular, and its practice spread across Europe and the United States. A large number of physicians also began to adopt the practice. The problem was that it didn't work, and as evidence against the method accumulated, phrenology began to lose its luster.

However, some of the assumptions that inspired phrenology have turned out to be true. In particular, the idea that different parts of the brain perform different functions is a central tenet of modern neuroscience. Unfortunately, many scientists were skeptical of that hypothesis because they associated it with a discredited pseudoscience. And perhaps as a result, the idea that specific areas of the brain might be particularly important for language took a while to gain acceptance.

Bouillaud and Auburtin

A couple of French physicians, Jean-Baptiste Bouillaud and his son-in-law Ernest Auburtin, firmly believed that different parts of the brain performed quite different functions. Based on the behavior of patients with brain damage, they argued that spoken language was particularly dependent on anterior parts of the left hemisphere.

Auburtin provided some pretty convincing evidence in support of their hypothesis. In particular, he came across a patient who tried to commit suicide by shooting himself in the head. But instead of killing himself, the man blew off only a part of his skull, exposing his brain, which was not actually damaged. And the exposed part of the brain happened to be in the anterior part of his left hemisphere. When Auburtin applied gentle pressure on that area, the patient's speech was suspended. It returned after the pressure was lifted.

Bouillaud and Auburtin suggested that the left frontal lobe is more involved in speech than in many other behavioral functions.

Bouillaud and Auburtin's ideas were met with considerable skepticism. But then along came a patient who would turn out to be one of the most important neurological patients in the history of neuroscience.

Expressive Aphasia and Broca's Area

Louis Victor Leborgne was born in France in 1809. He worked in the shoemaking business in Paris until around the age of 30, when he lost the ability to speak. The only thing he could say was a single syllable, *tan*, which he usually repeated as a pair, saying "tan tan."

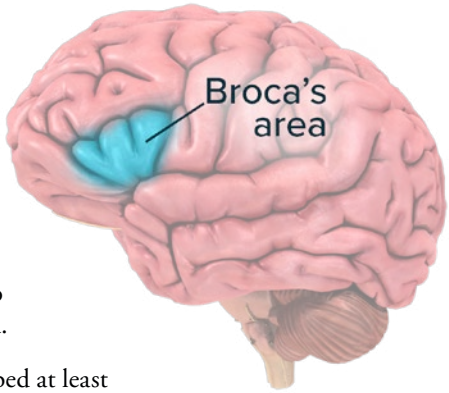
Leborgne behaved as if he knew what he wanted to say. He used hand gestures and seemed to understand what others said to him. A few years later, his physical health began to deteriorate. His vision and mental abilities also

began to decline, and he ultimately became bedridden. He subsequently developed gangrene, and that's when he met the French physician and surgeon Paul Broca.

Broca specialized in language, and he tested Leborgne extensively. He became convinced that Leborgne had a specific and selective deficit in speech production. Today, neurologists often refer to Leborgne's problem as Broca's aphasia. Another common name is expressive aphasia, to convey the fact that the problem with spoken expression is much more severe than comprehension.

Leborgne died in 1861, and his autopsy revealed brain damage in the left frontal lobe, particularly in the left inferior frontal gyrus. That same year, Broca presented his findings to the Anatomical Society of Paris and the Anthropological Society, and the scientific establishment became much more accepting of the idea that speech production might depend critically on the left frontal lobe. More generally, the hypothesis that different functions were localized in different brain regions began to be taken much more seriously.

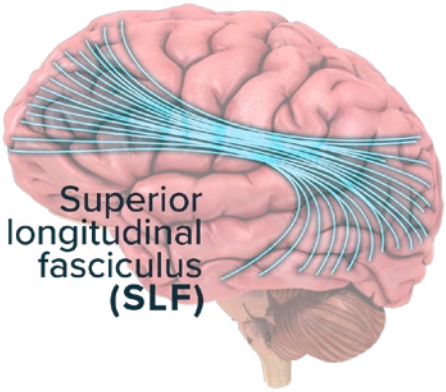
A few months later, Broca came across another patient, Lazare Lelong, who was also experiencing selective deficits in speech production. After Lelong's death, Broca found a lesion in his brain in the same place as the lesion in Leborgne's brain. Today neuroscientists refer to this part of the brain as Broca's area.



Over the next 2 years, Broca described at least 12 other patients with expressive aphasia and damage to the same area. His cases made it clear that this brain region is critical for speech production but plays a much less important role in language comprehension and most other mental abilities. The take-home message is that injury to one brain region can undermine one cognitive function without necessarily undermining others.

Both Leborgne's and Lelong's brains were preserved, and many scientists have traveled to Paris to study them. In one very famous study, Nina Dronkers collaborated with several French colleagues to conduct a high-resolution MRI study of the preserved brains.

They found that both brains did indeed exhibit damage in Broca's area, but they both also had significant damage in a major white matter tract called the superior longitudinal fasciculus, or SLF. The SLF is the major information highway connecting the front of the brain to the back of the brain. It allows frontal regions, like Broca's area, to communicate with regions in the back of the brain that are more important in perceptual processing, including perceiving and understanding language. The bottom line is that although Broca's area is indeed critically important for spoken language production, it's not the only brain region that plays a role.

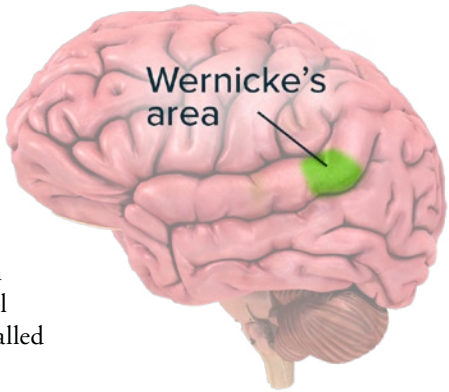


Receptive Aphasia and Wernicke's Area

A little over a decade after Broca described expressive aphasia, a German neuropsychiatrist named Carl Wernicke encountered another patient with language problems. While this patient had no trouble producing a fluent string of words, the words didn't make much sense. They didn't seem to go together or convey a coherent thought.

Furthermore, the patient didn't seem to understand spoken language. Wernicke referred to this type of aphasia—that is, a deficit in spoken language—as sensory aphasia to distinguish it from the kind of aphasia that Broca's patients exhibited. Today, it's typically referred to as Wernicke's aphasia or receptive aphasia.

When Wernicke's patient died, he discovered that the brain damage was in a completely different region than the damage in Broca's patients. Although it was still in the left hemisphere, it was much more posterior than the damage in Broca's patients. Specifically, it was near the junction of the temporal lobe and the parietal lobe, in a region that is now often called Wernicke's area.



So damage to Wernicke's area and damage to Broca's area produced a different pattern of deficits. Based on these differences, Wernicke hypothesized that the damaged region in his patient served as a repository for what he called word images, a term that refers to memories of how words sound. In addition, he hypothesized that Broca's area was critical for the production of word sounds.

Furthermore, Wernicke hypothesized that damage to the white matter fibers connecting the two regions would also produce aphasia but that the symptoms would be distinct from both Wernicke's aphasia and Broca's aphasia. His prediction turned out to be true: Patients with damage to the fibers connecting the two regions often exhibit what is now called conduction aphasia. Patients with conduction aphasia can understand what they are hearing but have difficulty repeating it back. Furthermore, they recognize mistakes in their own speech, but they have trouble correcting them.

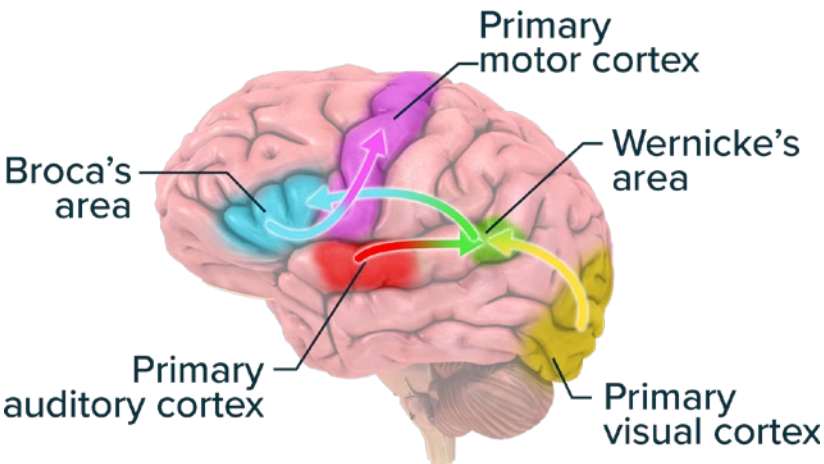
Wernicke associated the area named for him with language comprehension and Broca's area with language production.

Wernicke–Geschwind Model

Wernicke's model of the neural basis of language was revived and expanded by the famous neurologist Norman Geschwind in the late 1960s. He argued for a relatively simple model of language processing. Geschwind said that when someone is listening to spoken language, the auditory speech input is first processed by the primary auditory cortex in the superior temporal gyrus. The auditory information is then sent to Wernicke's area, where the words can be understood. If the person is reading instead of listening to spoken language, the information comes from the visual cortex but still ends up in Wernicke's area for language comprehension.

For language production, the words that the person wants to say are sent from Wernicke's area to Broca's area via underlying white matter tracts, and then Broca's area constructs the motor plans needed to say the words. Those motor plans are sent from Broca's area to the motor cortex, which is responsible for moving the mouth and tongue to actually say the words.

Using this very simple model, Geschwind was able to explain the behavior of many different neurological patients based on the site of their brain damage. For example, if the connections from the auditory cortex to Wernicke's area are damaged, patients can hear and speak, and they can understand written language, but they have a severe deficit in spoken language comprehension.



This makes sense in the Wernicke–Geschwind model because visual information can get to Wernicke's area, so reading is intact, but auditory information can't, and so there's no way to understand spoken language.

The Wernicke–Geschwind model was also able to make sense of a syndrome known as pure alexia. Patients with this problem can understand and produce spoken language but have a severe deficit in reading even though they can still see and recognize other visual objects. The Wernicke–Geschwind model naturally predicts this set of symptoms if the visual cortex gets disconnected from Wernicke's area but the auditory cortex remains connected. And the brain damage in these patients does seem consistent with that explanation.

Geschwind's work had a profound impact on neurology—especially his proposal that many neurological syndromes can be explained in terms of disconnections between specific regions with specialized functions. This proposal would become the dominant paradigm for understanding the behavior of brain-damaged patients and brain–behavior relationships more generally.

The Dual-Stream Model

The Wernicke–Geschwind model is still seen as a major landmark that shaped the field and provided a very helpful theoretical framework. But over the last 25 years, modern cognitive scientists have identified and attempted to address several important flaws.

For one thing, the original model doesn't consider the fact that language is hierarchical. For example, people produce and comprehend sentences. Those sentences are composed of phrases, which are composed of words. And words are often composed of subparts like prefixes and suffixes. Furthermore, each part of a word is composed of elementary language sounds called phonemes.

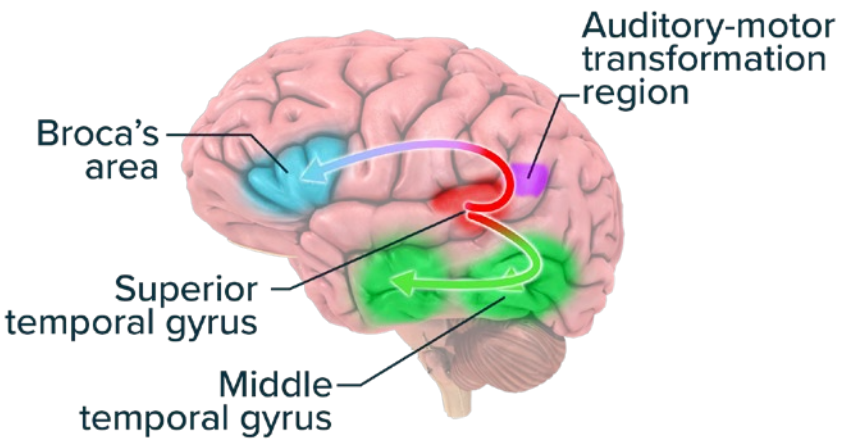
Aphasic patients can have problems at many different levels of this hierarchy. Unfortunately, the Wernicke–Geschwind model only deals with language at the level of word processing, and so it can't really explain a range of different symptoms.

Another problem with the model is that it makes some dubious assumptions about the key brain areas involved. Substantial evidence now exists that the areas involved in language processing extend well beyond those proposed by Wernicke and Geschwind and that the neurobiology of language is significantly more complicated than they assumed.

One very influential modern theory was developed by Gregory Hickok and David Poeppel—the dual-stream model of language processing, which is motivated by the intuition that people need to process speech sounds in at least two very different ways, and it assumes that those processes depend on different neural circuits.

First, people need to translate sounds, or phonemes, into meaning. Second, they need to map the same sequence of phonemes into a motor-based articulatory representation so that they can say the word.

The dual-stream model assumes that the processing stream that computes meanings starts in the superior temporal gyrus in both hemispheres and then sends information to the middle temporal gyrus. Conversely, the processing stream for saying the words also starts in the superior temporal gyrus but then sends the information somewhere else entirely. Information has to travel up and back to an auditory-motor transformation region near the junction of the temporal and parietal cortexes in the left hemisphere and then forward to Broca's area and nearby regions.



The dual-stream model suggests that one stream of processing deals with meaning and that the other deals with forming and saying the word aloud.

The dual-stream model represents a real advance over the classic Wernicke–Geschwind model. It incorporates a much larger set of brain regions, all of which have now been clearly implicated in language processing. It also more clearly describes the specific functions of those brain regions. And those functional assumptions match up much better with the pattern of deficits observed in brain-damaged patients than the assumptions of the Wernicke–Geschwind model.

Furthermore, the dual-stream model incorporates assumptions about language processing at different levels in the hierarchy, including phonological processing, articulatory processing, processing at the level of words, and combinatorial processing of words into phrases and sentences.

Reading

Brennan, J. R. *Language and the Brain: A Slim Guide to Neurolinguistics*. Oxford: Oxford University Press, 2022.

Friederici, A. D. *Language in Our Brain: The Origins of a Uniquely Human Capacity*. Cambridge, MA: MIT Press, 2017.

Obler, L. K., and K. Gjerlow. *Language and the Brain*. New York: Cambridge University Press, 1999.

5

Artificial Neural Networks and Language

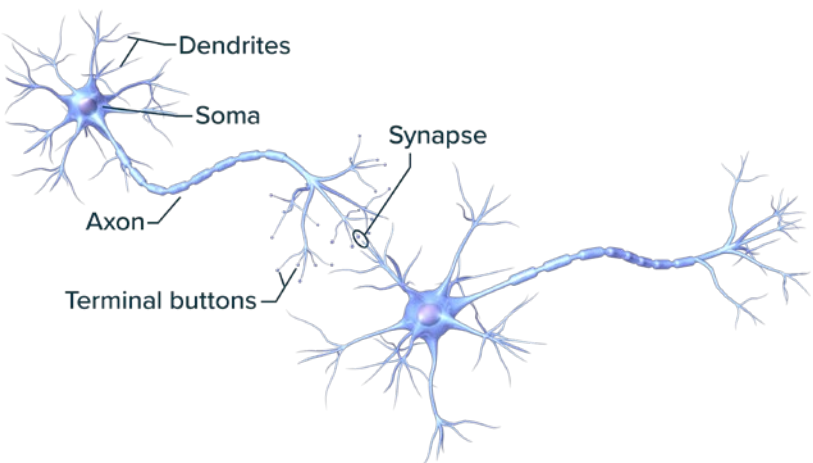
ChatGPT and other similar GPT models can write about anything imaginable in coherent, intelligible English. And in most cases, the sentences that they produce are difficult to distinguish from those produced by a real human being. How do they do it? This lecture aims to answer that question by taking you into the world of neural networks. You'll learn about what they are, how they work, and how they learn. You'll also learn about computational models of language and how far scientists have gotten in trying to implement language processing in computers.

Real Neural Networks

GPT models are deep artificial neural networks. These types of networks simulate some aspects of the real neural networks in the human brain. They involve networks of artificial neurons that are connected to each other by artificial synapses, which the neurons use to send signals to each other.

Real neurons have three main parts: the soma, the dendrites, and the axon. The soma is the cell body that contains the nucleus and all the genetic information. The dendrites are small branches that extend out of the soma in all directions. You can think of them as the input processes for the neuron. Most signals from other neurons come in via the dendrites. Finally, the long axon is the output process for the neuron. When the neuron wants to send signals to other neurons, it sends electrical impulses down the axon.

At the end of the axon are terminal buttons that often connect to the dendrites of another neuron at a synapse. The electrical pulses from the axon get communicated to other neurons via the synaptic connections.

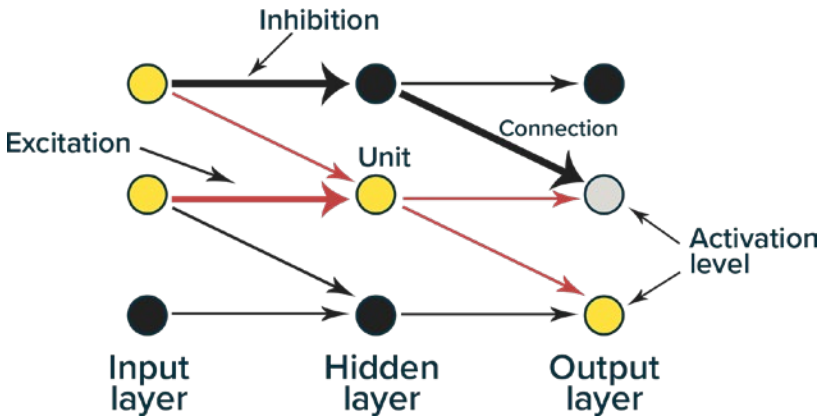


These connections between neurons can be either excitatory or inhibitory. If a connection is excitatory, then signals sent by the presynaptic neuron tend to excite the receiving postsynaptic neuron and increase its own chances of firing. Conversely, if the connection is inhibitory, then signals from the presynaptic neuron tend to inhibit the postsynaptic neuron and decrease its chances of firing. What determines whether the postsynaptic neuron will fire is whether it gets more excitatory or inhibitory input.

And this is how information gets sent around the brain: Some neurons fire, and they in turn cause some other neurons to fire, and so on. And that's what artificial neural networks try to emulate.

How Artificial Neural Networks Work

To get a sense of how artificial neural networks work, consider this simple model. Each of the circles is an artificial neuron or unit. Each of the arrows is an artificial synaptic connection. Each artificial neuron has an associated activation level, which is just a real number indicating how rapidly the unit is firing. The yellow units are firing a lot and have a high activation level, while the black units aren't firing very much and have a low activation level.



The artificial neuron units are organized in three layers from left to right. On the left, the input units represent different inputs using different patterns of activation. Meanwhile, the output units on the right represent the response to the input. The circles in the middle are called the hidden units because they're hidden between the input and output units and don't interact with the world outside.

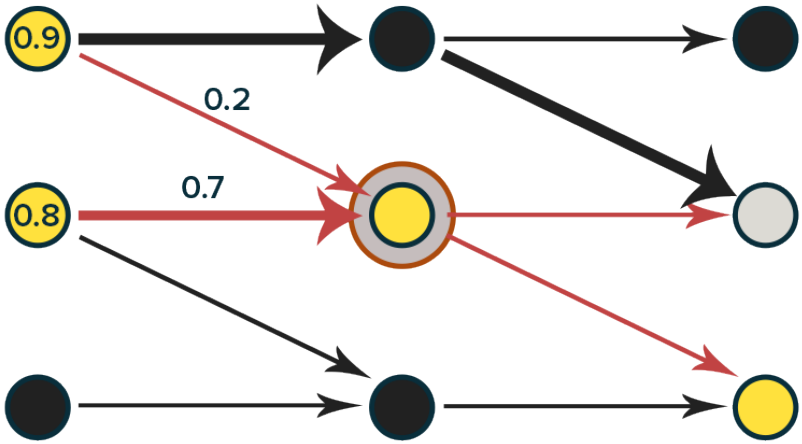
So if you wanted the neural network to do something, you would present an input pattern, meaning that you would set the activation levels for each of the units on the left. You'd set how fast each unit is firing, which equates to a specific firing pattern. That's the input. Once the network processes the input, the output will correspond to the pattern of activation across the output units.

As long as you come up with a standardized way to represent your inputs and outputs as numbers, then you can represent them in your neural network. The neural network can then map from every input pattern to an appropriate output pattern. The possibilities are truly limitless.

To map from input patterns to output patterns, the neural network needs to send signals from the inputs through the hidden units to the output units. And to do that, it uses artificial synapses.

Each of these artificial synaptic connections also has a number associated with it. But in this case, that number represents two things: whether the connection between units is excitatory or inhibitory and the weight or strength of the connection. In the graphic, excitatory connections are red, and they correspond to positive numbers. Inhibitory connections are black, and they correspond to negative numbers. Stronger connections are indicated by thicker arrows and correspond to bigger numbers.

So, how might the artificial neuron in the center of the following graphic simulate neural information processing? A few numbers have been added so that you can walk through a basic computation. For simplicity, assume that activation levels can go from 0 to 1, where 0 means the unit is not firing at all, and 1 means the unit is firing at its maximal firing rate. Likewise, assume that connection strengths can go from –1 to +1, where –1 means the strongest possible inhibitory connection, and +1 means the strongest possible excitatory connection.



The bright yellow unit in the top left has an activation level of 0.9, which is close to 1, so that unit is very active. Likewise, the unit just below it has an activation level of 0.8, which is also quite active.

All the artificial synaptic connections, which are represented by the arrows in the graphic, also have numbers associated with them. Those numbers correspond to the strength of the connection and whether it's excitatory or inhibitory.

The connection between the top unit on the left and the middle unit in the center is colored red, which tells you that it's excitatory, but the arrow is quite thin, so it's a relatively weak connection. Accordingly, the number associated with that connection is 0.2, which is a positive but small number. It's a weak excitatory connection.

The connection below that has a weight of 0.7, which indicates that the connection is excitatory. And the number is closer to 1, which means it's a relatively strong excitatory connection. So the connection is colored red because it's excitatory, and it's also relatively thick, indicating that the connection is strong.

As for the unit in the middle of the network, you don't know its activation level, but you can figure it out. First, you have to figure out how much input the unit is getting. Then you have to translate that input into a specific activation level.

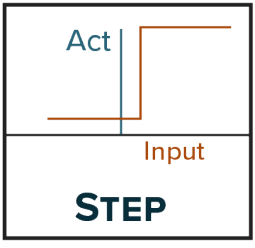
First you use the excitatory input from the two units on the left to compute the total amount of input to the middle unit. So you add up the input coming from each of the presynaptic units. And the input from each unit is just the activation level of that unit multiplied by the connection between it and the postsynaptic unit.

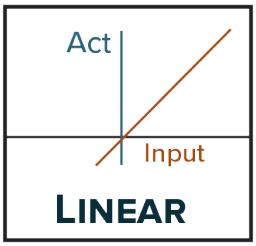
Consider the unit on the top left. Its activation value is 0.9, and it has a weak excitatory connection with a value of 0.2. If you multiply these two numbers, you get 0.18, so that's the input coming from the top-left unit.

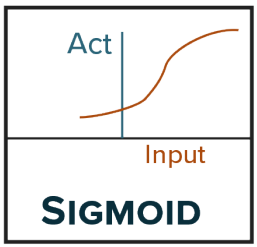
Now consider the input coming from the middle unit on the left. Its activation value is 0.8, and the strength of its connection to the unit in the middle is 0.7. So the input coming from that unit is 0.8 multiplied by 0.7, which is 0.56. When you add the two inputs, you get 0.74 as the total input to that unit.

$$\begin{aligned}\text{Weighted input} &= (0.9)(0.2) + (0.8)(0.7)\\ &= 0.18 + 0.56\\ &= 0.74\end{aligned}$$

Next you need to translate the total input into a specific activation level. In other words, you need to figure out how much the unit will fire when it receives that input. And a number of ways to do that exist. Some neural networks use binary units that are either fully on or fully off. If the total input is above some threshold, such as 0.5, then the unit's activation would be 1. Otherwise, it would be 0. That's sometimes called a step function.





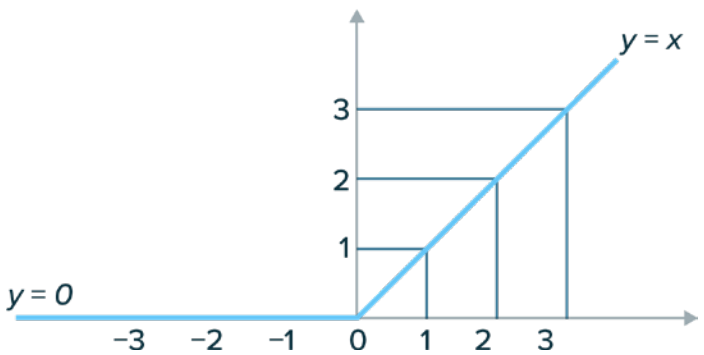


Another approach would be to use a linear activation function, meaning that the activation level always goes up by the same fixed amount with each unit increase in input.

One of the most common activation functions is the sigmoid function, which looks like an *S*. The nice thing about the sigmoid function is that activation levels continuously increase as the input to the unit increases, but there's also an upper and lower bound on activations. So if you always want your activation levels to vary continuously between 0 and 1, then you might want to use a sigmoid activation function.

Many modern artificial neural networks, including GPT systems, use rectified linear units. These units simply use their input as the activation level unless the input is negative, in which case they produce an activation level of 0. This type of activation function is popular because the computations involved are simple and fast, which helps when you are building and training very large networks in which efficiency is an important consideration.

RECTIFIED LINEAR UNITS



Now you understand how an individual unit in this neural network functions. And of course, that's what all the units throughout the network are doing all the time. Each one is computing its total net input and then running that through an activation function to update its activation level. Every new activation level then gets passed on as input to other units downstream, once again weighted by the strength of the connections between the units. That's how information processing happens in artificial neural networks like ChatGPT and how GPT systems can answer questions and produce humanlike language output.

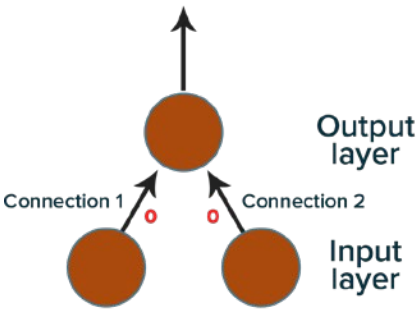
Supervised Learning and Backpropagation

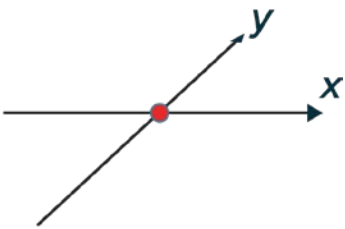
The network requires a lot of training to produce the output that you want it to produce. Basically, you repeatedly give it a pattern of activation across the input units, and you also give it an output pattern that you want it to produce when given that input. Before you train it, the neural network will produce random outputs. But if you compare the output that it actually produces to the output that you want it to produce, then you can measure the difference between them and make small adjustments to the weights of the connections so that the network will produce an output that's closer to the target next time. And if you do this over and over again, then the network eventually learns to produce the targets you want it to produce.

This approach is called supervised learning because you're constantly telling the network what you want it to do, and the network is repeatedly adjusting its weights to get closer and closer to what you want. The real trick is figuring out how to adjust the weights. The most common solution is an algorithm called backpropagation, which was one of the most important developments in the history of neural network research.

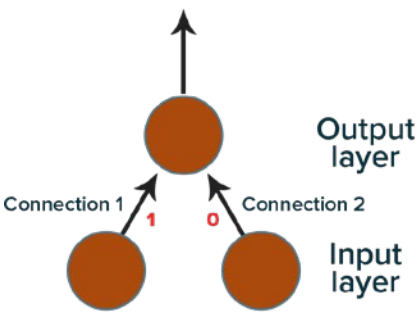
The backpropagation algorithm uses a technique called gradient descent. Imagine that you have a very simple neural network that has only two input units. Both of these input units are directly connected to one output unit, with no hidden units at all.

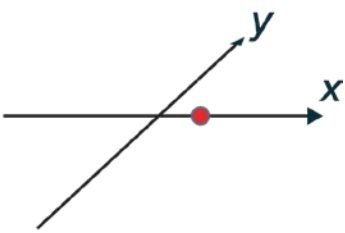
The strength of those two connections is represented on a horizontal *x-y* plane. You can plot the strength of connection 1 from left to right and the strength of connection 2 from front to back. Then any pair of connection strengths can be represented by a single point on this plane. If both connection strengths are 0, then you're at point 0,0.



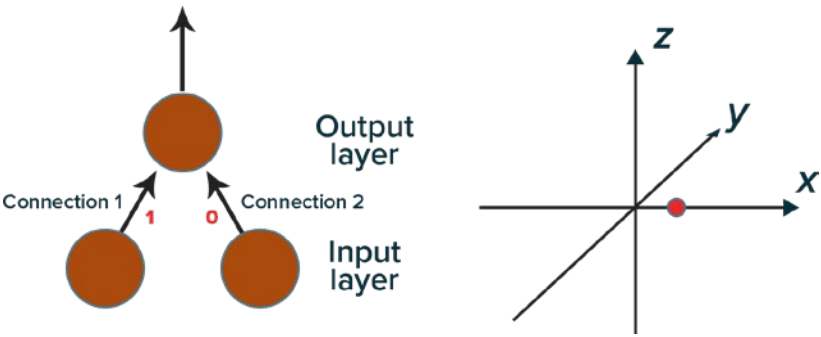


If connection 1 has a strength of 1 while connection 2 has a strength of 0, then you're at point 1,0.





A third, vertical dimension can be added to represent how big of a difference exists between the output pattern that the network actually produces and the target output pattern that you want it to produce. You can call that difference the error. The bigger the error, the higher the point on the surface.

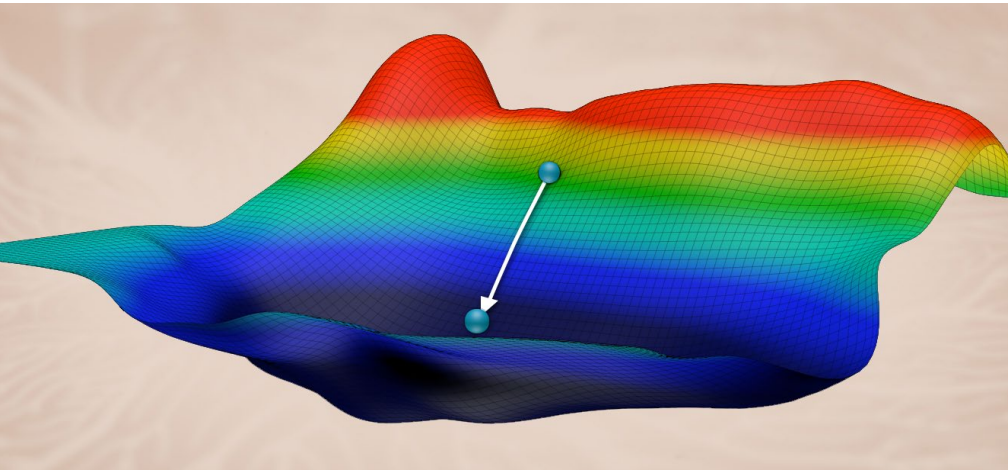


In this case, you only have a single output unit. So the actual output and target output will both be single numbers. You can subtract one from the other, and the difference tells you the error. And you usually want your error measure to be positive, so you typically just square the difference between the numbers.

Now you can plot the squared difference between the actual output and the desired or target output as a function of the connection weights. Every point on the horizontal plane corresponds to a pair of connection weights, and for every pair, you're plotting how much error in the output that pair of connection weights produces.

You want to find a pair of connection strengths that produces the smallest error, so in the following figure, you want to try to find the minimum. The minimum—the smallest error you can find—will always sit at the lowest value on the surface. If you could see the whole surface, then you could quickly find the minimum, but you can't see the whole surface. You have only local information about the current set of weights.

So you want to make a small change in the weights that reduces the error. In this figure, that might correspond to moving the dot, which represents your current set of weights, a small distance on the *x*-*y* plane such that the error is reduced.



And that's where the gradient comes in. The gradient tells you the slope of the surface in every direction. It uses calculus to tell you the best direction to move the black dot: basically, the direction where the steepest slope is going down. Imagine that you're on top of a mountain and you want to get to the bottom. You look around and take a step in the direction of the steepest downward slope. Then you look around again and take a step in the direction of the next steepest downward slope. And so on. That's the idea of steepest gradient descent.

And that's what the backpropagation algorithm does. It repeatedly computes gradients based on the current connection weights and then changes the weights ever so slightly in a way that reduces the error the fastest. And it does this over and over again for hundreds, thousands, or even millions of trials until it finds a set of connection weights that produces the desired mapping from inputs to outputs with very little error.

How GPT Systems Produce Language

The task that GPT systems perform is simple: Given a sequence of words, predict the next word. For example, if you hear someone say, "I'm low on funds. I'm going to stop by the ATM and get some … ," you might predict that the next word would be *cash* or *money*.

The way to get the GPT system to answer questions in complete sentences or to write stories is to feed back the word it just produced and add it as the last word in the input, removing the first word from the previous input. Then you ask it to predict the next word again. You add that word to the end of the input and ask it to predict the next word again. And it turns out that GPT systems are extremely good at predicting words that make sense given the previous context. But that takes a lot of training.

GPT systems also incorporate a few other key ideas that have proven to be very powerful. One is that they can process thousands of tokens of text as input at a time rather than presenting a single word as input. In this way, they can take into account words that were presented quite a while ago, which can impact what the next word will likely be.

GPT systems also incorporate positional information about words, which allows the neural network to distinguish different word orders. That's important because, for example, the sentence *Mary doesn't like John* is very different from the sentence *John doesn't like Mary*, even though they consist of exactly the same words.

A third key idea is known as attention. GPT systems include connections between all pairs of tokens among the thousands of input tokens. The attention mechanism allows the systems to learn about such long-range associations between any pair of tokens in the long input stream.

Transformer Networks

Artificial neural networks that are trained to predict the next word, that are given a large number of input tokens simultaneously along with positional information, and that use an attention mechanism to learn associations between pairs of input tokens are called transformer networks or, more

specifically, generative pretrained transformer networks—thus the initialism GPT. They're also often referred to as large language models because of the large amount of data they are trained on.

These systems behave much more intelligently than other computer programs ever have, and that's why they're one of the most important developments in the history of artificial intelligence. They can write fictional stories that sound real, provide plausible answers to questions on virtually any topic, and translate between any pair of common natural languages. The list goes on.

Proponents would argue that transformer networks represent a major step forward in efforts to build an artificial general intelligence. But not everyone agrees. One of the main counterarguments is that the ability to predict the next word is very different from true understanding. The claim is that GPT systems don't really understand what they're talking about. They just spit out the most likely next word based on analyzing what typically comes next on the internet. And that can lead to behavior that doesn't seem particularly reliable or intelligent.

These systems do extremely well, but they're not perfect. They make mistakes and sometimes make statements that are demonstrably false, which is consistent with the hypothesis that being able to predict the next word doesn't constitute true human understanding. Nevertheless, they certainly represent a very interesting model of language that every student of cognitive science should know about.

Reading

Aggarwal, C. C. *Neural Networks and Deep Learning: A Textbook*. 2nd ed. Cham, Switzerland: Springer Nature Switzerland, 2023.

Bengio, Y. "Machines Who Learn." *Scientific American* 314, no. 6 (2016): 46–51.

Wolfram, S. *What Is ChatGPT Doing … and Why Does It Work?* Champaign, IL: Wolfram Media, Inc., 2023.

6

How Babies Think about the World

Human beings spend more time being dependent on their parents than just about any other species. Why? One popular theory is that children need a lot of time to learn the behaviors and acquire the abilities that will allow them to be adaptive and to thrive in whatever environment they find themselves. From this point of view, what babies and children excel at is learning. This lecture explores what babies seem to know about the world in terms of physical, psychological, and moral reasoning.

Physical Reasoning

A major methodological challenge exists associated with studying babies and how scientists overcome it. Babies obviously have only rudimentary language skills, so you can't ask them to do the same experimental tasks that adults perform. Furthermore, you can't ask babies what they're thinking about or how they're performing a task. So how can you learn anything about their internal mental life?

Infants are like adults in that they can get surprised by what they see. And also like adults, infants spend more time looking at things that surprise them. By carefully analyzing where infants look and what they look at longest, researchers can learn what infants find surprising and, conversely, what they expect to see. The idea is that if scientists can figure out what babies expect to see, they can figure out what babies know about the world. And this kind of violation-of-expectation paradigm has led to several new insights into infant cognition.

For example, until the 1980s, most developmental psychologists believed that babies less than 8 months of age did not understand that objects continue to exist when they go out of sight. This hypothesis was supported by the research of the famous Swiss developmental psychologist Jean Piaget, who argued that young babies lack a sense of object permanence and therefore believe that hidden objects no longer exist.

However, beginning in the 1980s, studies using the violation-of-expectation paradigm found compelling evidence that babies much younger than 8 months old do have a sense of object permanence. In one of the earliest studies, Renee Baillargeon, Elizabeth Spelke, and Stanley Wasserman presented 5-month-old infants with two different events: a possible event that was consistent with object permanence and an impossible event that violated object permanence. In other words, one scenario played out how one would expect it to in real life, where objects remain even though

Babies and very young children understand that physical objects persist in space and time, and they even have a rudimentary understanding of motion, inertia, and gravity.

they disappear from sight. The other scenario played out as if hidden objects ceased to exist. The researchers measured how long the babies looked at both types of scenarios.

They found that the babies looked at the impossible event significantly longer than they looked at the possible event, meaning they were surprised by the outcome of the impossible event. Subsequent violation-of-expectation experiments have confirmed this finding and extended it in many ways.

Other experiments have demonstrated that babies also have different expectations for inert objects versus self-propelled objects. In particular, if an object appears to be inert, like a rock, then they're surprised if it suddenly begins to move on its own. They're also surprised if it autonomously changes direction after being set in motion or if it doesn't fall when released above the ground. In contrast, young babies aren't surprised by any of these outcomes if the object appears to be self-propelled, like an animal. So apparently, very young babies know that self-propelled objects are fundamentally different from inert objects and that they can control their own motion even in the face of external forces.

Psychological Reasoning

Several studies have investigated the psychological reasoning of babies—what they know about other people's thoughts and behavior—and at the most general level, the findings indicate that babies typically expect people to behave rationally. In one of the earliest studies, György Gergely and his colleagues at the Hungarian Academy of Sciences in Budapest showed 1-year-old babies a movie of an animated ball that had to jump over a wall to reach another animated ball that it wanted to be next to.

Babies appreciate that other people have goals in mind, and they expect people to behave in ways that help them achieve their goals efficiently.

After repeatedly showing this movie, the scientists then presented the babies with two different movies. They showed the same balls, but the wall was removed. In one movie, the ball jumped just as it had before. In the other movie, the ball just headed straight toward the other ball without jumping.

The babies looked longer at the movie in which the ball jumped instead of heading straight for the other ball. Apparently, once the wall was removed, the babies expected the ball to head directly to the other ball and were surprised when it jumped.

These results provide compelling evidence that 1-year-olds possess what's sometimes called a theory of mind. That is, they understand that some other things in the world have a mind and that those things can have goals that they are trying to achieve. Furthermore, babies understand that those things will try to achieve their goals in a simple, efficient way.

There's also evidence that impairments in developing a theory of mind might be an important symptom of autism. In one famous experiment, Simon Baron-Cohen, Alan Leslie, and Uta Frith at the MRC Cognitive Development Unit in London recruited three groups of young children: one group with Down's syndrome, one group with autism spectrum disorder, and one control group that did not exhibit either of these disorders. They asked all the children to perform the so-called Sally–Anne test.

In this test, the children saw two girls, Sally and Anne, in a room. They saw Sally hide a marble inside a covered basket and then leave the room. While Sally was away, Anne took the marble out of Sally's basket and put it in a box. Sally then returned to the room.

The children were asked three questions:

1. Where was the marble in the beginning?
2. Where is the marble really?
3. Where will Sally look for her marble?

The scientists found that all the children answered the first two questions correctly. But when asked the third question, the children with autism spectrum disorder answered very differently from the other groups. For example, 85% of both the control children and the children with Down's syndrome thought that Sally would look in her basket rather than in the box. Even though they knew the marble was no longer in the basket, they also realized that Sally didn't know that. And by thinking about the situation from Sally's perspective, they predicted that Sally would look in the wrong place.

In contrast, 80% of the children with autism spectrum disorder thought that Sally would look in the box rather than in the basket. So even though they knew where the marble had been at the beginning and where it was now, they predicted that Sally would look where the marble actually was, without realizing that Sally wouldn't know that it had moved.

The scientists argued that the children with autism spectrum disorder had an impairment in their theory of mind and therefore had difficulty thinking about the situation from Sally's perspective. The jury is still out on the idea that a deficit in theory of mind plays a central role in autism, but the idea has generated an enormous number of studies that are continuing to be done today.

Moral Reasoning

J. Kiley Hamlin, Karen Wynn, and Paul Bloom investigated whether babies have a sense of right and wrong. Their influential study was conducted at Yale University. In it, 6- and 10-month-old babies saw a red wooden circle with eyes that was trying to climb a hill, but it wasn't able to make it to the top on its own. The babies repeatedly saw a yellow triangle come along and help the red circle get to the top by pushing it from behind. They also repeatedly saw a blue square hinder the red circle by blocking its path and pushing it down the hill. Later, when the yellow triangle and the blue square were both placed within the babies' reach, both groups of babies tended to reach for the shape that helped rather than the shape that hindered the red circle.

Subsequent experiments also found that infants preferred helpers over neutral agents and neutral agents over hinderers. These results have now been extended to 3-month-olds, suggesting that some of these preferences may be innate.

Other studies revealed that these preferences were observed only if the original agent was presented as being alive. If the babies saw the same scenarios but with inanimate objects, they no longer exhibited the preference. In other words, the babies preferred actions that helped an agent they perceived to be alive, which sounds a lot like a moral preference.

So how do babies respond when they see agents physically harm others? Yasuhiro Kanakogi and several of his colleagues examined this question in an experiment they conducted at Kyoto University in Japan. They presented 10-month-olds with a colored shape that acted aggressively toward a victim shape, running into the other shape and pushing it into a wall. When the shapes were placed within the babies' reach, the babies tended to reach for the victim shape and avoided reaching for the aggressive shape. They also added a neutral shape that moved around but didn't interact with the aggressive or victim shapes. In that case, they found that the babies reached for the victim shape significantly more than the neutral shape, which the authors interpreted as evidence of sympathy.

Children appear to have a sense of right and wrong from a very young age. Babies are sensitive to differences between helping and harming, between victims and aggressors, and between fair and unfair behavior. Studies suggest that babies are also able to distinguish between prosocial and antisocial behavior.

Evidence also suggests that babies can distinguish fair behavior from unfair behavior. For example, Marco Schmidt and Jessica Sommerville showed 15-month-old babies movies in which a person shared four graham crackers with two other people. In the "fair" movie, both people got two crackers, but in the "unfair" movie, one person got three and the other person only got one.

The scientists found that the babies looked significantly longer at the unfair outcome than at the fair outcome, suggesting that they were sensitive to the difference and surprised by the unfair outcome. When the scientists showed the babies the same outcomes but without any people in the situation, the babies apparently no longer perceived the unequal distribution as unfair since no people were involved.

Reading

Bloom, P. *Just Babies: The Origins of Good and Evil*. New York: Crown, 2013.

Gopnik A. "How Babies Think." *Scientific American* 303, no. 1 (2010): 76–81. https://doi.org/10.1038/scientificamerican0710-76.

Siegal, M. *Marvelous Minds: The Discovery of What Children Know*. Oxford: Oxford University Press, 2008.

Spelke, E. S. *What Babies Know: Core Knowledge and Composition*. New York: Oxford University Press, 2022.

7

Working Memory: The Mind's Notepad

Just about any demanding cognitive task requires you to use your working memory, whether you're doing any kind of problem-solving or just trying to remember a list of words. In this lecture, you'll explore verbal and spatial working memory and see how cognitive scientists use artificial neural networks to simulate many aspects of working memory. You'll learn how the brain keeps information active even after the perceptual input has been removed and why that information might decay over time.

Verbal and Spatial Working Memory

Working memory refers to the ability to store information for a brief period of time in the service of your goals. For example, you use your working memory to keep track of what's been said in a conversation.

Likewise, when you're reading, you're constantly storing information and retrieving it later to relate it to new information.

But what kind of information do you store in working memory? For example, when trying to remember a list of words, do you store a representation of what the words mean or a representation of how the words are spelled?

A substantial amount of evidence now suggests that what you store is a representation of what the words sound like, or what cognitive scientists call phonology. Additional evidence indicates that you can refresh the information in your working memory by rehearsing it—saying it to yourself, either out loud or silently.

And it turns out that messing with your ability to rehearse can undermine your working memory because you rehearse items in working memory using the same speech mechanisms that you use to talk. When you talk out loud, you can't use those mechanisms to rehearse the items in working memory, and so they're harder to remember.

Most people can rehearse only about 2 seconds' worth of information, which means that people can remember more short words than they can long words. And people who talk faster can typically rehearse more words, and therefore store more words, than people who talk slower.

You also have a separate working memory system that stores spatial information. You use this system to briefly store locations. For example, suppose you're cooking a meal, and you've taken out a bunch of ingredients that you'll need and have put them on your kitchen counter. You would use your spatial working memory system to temporarily store those locations so that you can remember where your ingredients are a few seconds later.

Spatial working memory is independent of verbal working memory, and they use completely different neural circuits. But the two systems nevertheless share a number of similarities. For example, just as talking out loud interferes

with verbal working memory, requiring people to perform a spatial task, such as tracking a moving spot of light, interferes with spatial working memory. Furthermore, rehearsal seems to play an important role in refreshing information in both types of working memory.

Assumptions about Neural Computation

The goal of cognitive scientists is to develop computationally explicit theories about the detailed mechanisms involved in working memory. One of the questions they want to answer is: If neurons need some kind of input to maintain a high level of activity, then how does the brain maintain a representation of a word or location in working memory if the perceptual input is no longer present?

Cognitive scientists have discovered that by adopting three simple and widely accepted assumptions about how neural computation works, they can simulate many aspects of human working memory. Those three assumptions are distributed representation, recurrent connectivity, and Hebbian learning.

Distributed representation has to do with how people's brains represent information. Representations are generally distributed across a population of neurons rather than localized to individual cells. For example, when you're representing the sound of the word *dog*, hundreds of thousands of neurons fire to represent that sound. You don't just have a single "dog" neuron. Distributed representations are robust, meaning that the representation can easily survive the loss of a few neurons here and there without disrupting the representation.

With distributed representations, you also get automatic generalization. For example, imagine that you have a neural representation of a robin that is distributed over a population of hundreds of thousands of neurons. And those neurons have synaptic connections with hundreds of thousands of other neurons. So when the representation of a robin is active, it can then activate lots of other associated pieces of information, like the fact that robins can fly and that they lay eggs.

Now imagine that you perceive a starling or a sparrow when you're out walking one day. It's not a robin, but it obviously shares a lot of similarities with a robin, which means that the neural representation of this new bird will

overlap substantially with the neural representation of a robin. Although the activation patterns won't be identical, most of the same neurons will still be activated. The similarity in the distributed activation patterns allows you to generalize from other birds to this new bird you just encountered, and you'll associate it, too, with flying and laying eggs.

Recurrent connectivity refers to the fact that connections between different parts of the brain typically go in both directions rather than only one way. This recurrent connectivity is what allows neural networks to maintain activity even after input has been removed. The reason for this is that you can get cycles of activation in which neurons activate other neurons, which then activate other neurons, until eventually the original neurons get reactivated, starting the cycle again.

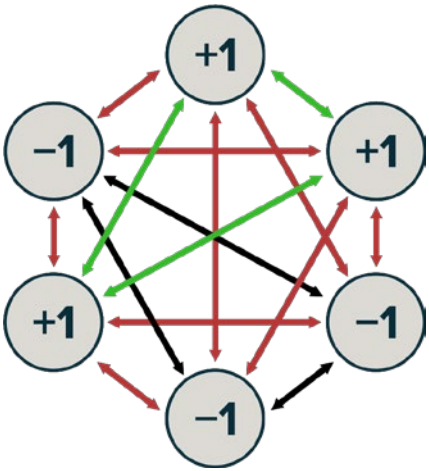
Recurrent connectivity can also produce the kind of time-based decay of information that is often seen in working memory. Specifically, if the recurrent input coming back to the original neurons isn't quite as strong with each passing cycle, then the distributed pattern of neural activity will get weaker and eventually die away completely.

Hebbian learning is often summarized with the phrase "Neurons that fire together wire together." Imagine a presynaptic neuron that's connected to a postsynaptic neuron by a synaptic connection. If both neurons are firing at the same time, then the synaptic connection between those neurons will be strengthened. This idea was first proposed by the famous Canadian neuropsychologist Donald Hebb.

Simulating Aspects of Working Memory

When you combine Hebbian learning with distributed representations and recurrent connectivity, the combination leads to some interesting computational dynamics that have proven very useful in explaining many aspects of working memory.

In this example of a very small artificial neural network, the six circles represent six artificial neurons. The arrows represent complete recurrent connectivity—every neuron is connected to every other neuron.



Across the population of neurons is a distributed pattern of activity. For simplicity, only two activation levels occur: +1 and –1. The neurons that are labeled with a +1 are firing a lot, while the neurons labeled with a –1 have a low firing rate.

Now apply Hebbian learning. You can assume that this distributed pattern of activation has appeared many times and that the synaptic connections between the units have been updated each time. According to the Hebbian rule, the connections between neurons that are both active get strengthened. So the connections between the units that are labeled +1 should be strong and excitatory. They are represented in the graphic by thick green arrows.

A common extension of Hebbian learning is to assume that if two neurons are connected but one is firing while the other is not, then the connection between them should become more inhibitory. The inhibitory connections are represented by red arrows.

Now consider the behavior of this basic neural network. The three neurons that are labeled +1 are all connected by strong excitatory connections, meaning that they will create a reverberating cycle of activity in which one unit excites another, which in turn excites another, and so on. And that activation will continue cycling in that circuit so that the active neurons can stay active even if there's no more external input coming in.

This critical mechanism allows neural networks to simulate working memory. For tasks that require working memory, you need to maintain information across brief delays, even after the original input has been removed—exactly the functionality that these kinds of neural networks provide.

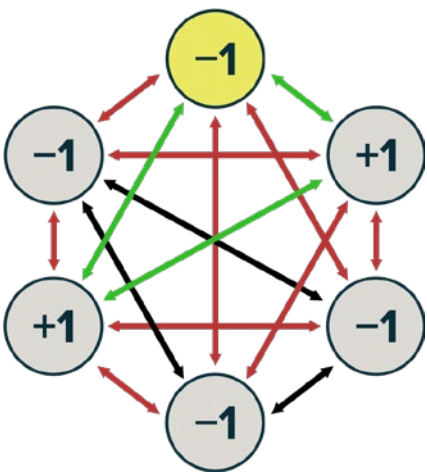
However, the neurons will receive less input once the external input is removed because the reverberating input has become the only source of input. As a result, the neurons may fire a little less and therefore provide less reverberating input to their neighbors. As time passes, the activation pattern will decay and ultimately disappear if it's not refreshed by additional external input. And in most computational models, rehearsal of the items provides that additional external input to refresh the items in memory.

Pattern Completion and Attractor Networks

The kinds of neural networks that have been discussed in this lecture have some other interesting properties that are also relevant for memory. In particular, they can perform pattern completion.

To illustrate the idea, consider again the simple recurrent network in which the connection strengths were set by Hebbian learning. But instead of the originally stored distributed activation pattern, you now have a noisy version of that pattern. Specifically, the value of the neuron at the top of the network has changed from +1 to –1.

Once the activity starts spreading around the network, the top neuron will be getting strong excitatory input from the neurons that are labeled +1. It also has inhibitory connections coming in from three other neurons. But those neurons are labeled –1, which means that they're not active, and so they won't pass along much input to that top neuron.



Because the top neuron is getting a lot of excitatory input, it's going to immediately become active again. And once it does, you're back to the original pattern that was stored, which will stick around but slowly decay. And the same is true for every other neuron in the network: If you change its activation value, all the other neurons in the network will immediately change it back.

The network is attracted to the stored pattern, and it will tend to return to that stored pattern anytime it's given an activation pattern that is similar. The stored pattern is therefore called an attractor for the network, and the network itself is typically called an attractor network.

If you set up the network so that it stores a particular image and give the network a degraded form of that image as input, it will quickly clean it up and return to the stored image. Likewise, if you give it part of the image as input, it will tend to complete the pattern and again return to the image that is stored in memory.

And this kind of functionality is directly relevant to memory. For example, suppose you want to remember that one particular face is associated with the name Bob and that another face is associated with the name Barb. If you present the attractor network with Bob's face, it will immediately retrieve the name Bob, and if you present it with the name Barb, it will retrieve the pattern corresponding to her face.

Attractor networks also provide a very natural explanation for why people have a hard time remembering similar items. It's much harder for an attractor network to store distributed patterns of activity that are similar to each other than to store patterns that are very different.

Attractor networks can store multiple different activation patterns across the same set of connections. Each of those patterns can be a separate attractor, each one representing different information. And the network will settle into the attractor pattern that is most similar to the pattern it starts in. But there are two conditions.

First, you have to have enough neurons so that a few bad connections won't cause problems because there will be a lot of other connections that are correct for that pattern. And those other connections will overwhelm the small number of connections that are trying to turn off a neuron that should be active or turn on a neuron that should be inactive.

Second, the attractor patterns that you want to store can't be too similar to each other. For example, you wouldn't be able to store two activation patterns that are identical except for neuron X being active in pattern 1 and inactive in pattern 2. You need a bunch of the other neurons to have different activation values so that the network can distinguish the two patterns and decide whether neuron X should be active or not.

And of course, this kind of problem is exactly what you encounter in working memory. When you want to store items that are very similar, your neural network runs into problems. You'd do much better if the items you're trying to store are very different from each other. Form this lecture, you should now understand the cognitive science behind this phenomenon.

Reading

Amit, D. J. *Modeling Brain Function: The World of Attractor Neural Networks*. New York: Cambridge University Press, 1989.

Baddeley, A. *Exploring Working Memory: Selected Works of Alan Baddeley*. London: Routledge, 2017.

Cowan, N. *Working Memory Capacity*. New York: Psychology Press, 2005.

8

Episodic Memory: A Library of Times and Places

Human beings are capable of storing enormous quantities of information and holding onto it for a long time. But memories can also become distorted to such a degree that people can misremember events that had a major impact on their lives. So how do scientists reconcile these seemingly contradictory aspects of memory? This lecture dives into this paradox and the mechanisms of episodic memory—that is, long-term memory for personal episodes from your own life. You'll learn how people store information about their personal past and why that information can sometimes decay.

Episodic Memory

Episodic memories are tied to a specific time and place and are remembered from a first-person perspective, as if you can mentally travel back in time to the original event. They can last for days, months, or even years. Your memory of attending a wedding, eating at a restaurant, or visiting a beautiful vacation destination would all be examples of episodic memories.

The first remarkable feature of episodic memory is its virtually limitless capacity. You're constantly laying down new episodic memories every minute of every day. And you're doing this throughout your entire life.

Of course, memories do decay and become less accessible over time if they're not rehearsed. But even very old memories that haven't been rehearsed in years can still be retrieved with the right cues. For example, if you've ever visited a place where you lived many years ago, you probably had the experience of remembering events that you hadn't thought about for years.

Within the past couple of decades, cognitive scientists have identified and studied a few people whose autobiographical memory is so good that it's kind of hard to believe. Consider the case of A. J., a woman who was tested by Elizabeth Parker, Larry Cahill, and Jim McGaugh at the University of California, Irvine.

A. J. seemed to be able to bring to mind every day of her life since she was 14 years old and tell you what happened to her that day as well as any significant historical events that occurred. She had kept detailed diaries throughout these periods of time, and even though she had no idea what dates the researchers were going to ask her about, the memories she produced invariably matched what she had written in her diary decades earlier. She also always got the days of the week right. And whenever she mentioned a historical event, such as Richard Nixon's death, she got that right, too.

Interestingly, A. J. is not unique. A few dozen other people with highly superior autobiographical memory have since been identified and tested. The researchers call this superior memory hyperthymesia, from the Greek words *thymesis*, meaning "remembering," and *hyper*, meaning "more than normal."

Obviously A. J. and people like her are out there on the extreme, but it turns out that the average person also has a pretty amazing memory, especially for visual information. In one famous experiment at Bishop's University in Canada, Lionel Standing asked normal people to look at thousands of pictures for 5 seconds each. They got to see each picture only once, and then 2 days later, they were asked to perform a forced-choice recognition test. During the test, they were presented with pairs of pictures, and each pair contained one picture they had seen and one they hadn't seen. They had to choose the picture they had seen before.

People were more than 95% accurate at remembering 1,200 striking pictures. And if people were tested immediately after the study instead of 2 days later, most of them didn't make any errors after studying 1,000 vivid pictures for 5 seconds each.

Memory Distortion and Decay

Although people can store huge quantities of information in their memories, that information can and does get distorted over time. People lose information as it decays, fill in gaps with plausible guesses that aren't always accurate, and modify memories to better fit with their experience and expectations. And most of the time, people are completely unaware that their memories are distorted.

Elizabeth Loftus at the University of California, Irvine, has done several studies that demonstrate the point. In one famous experiment, participants watched a videotape of a car accident, and then their memory of what they saw was tested. In the real videotape, there was no broken glass, but 14% of the people claimed that they remembered seeing broken glass.

Furthermore, Loftus found that she could subtly mislead people and make them more likely to remember seeing broken glass. She asked them, "How fast do you think the cars were going when they smashed into each other?" The simple phrase "smashed into" suggested that the collision was a violent one. And sure enough, 32% of people who were asked that question said that they thought they remembered seeing broken glass.

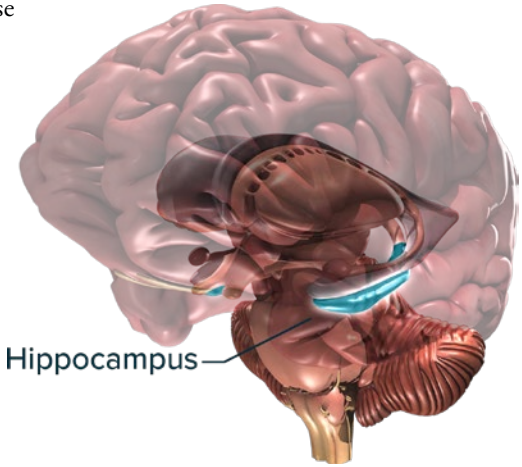
Everyone's memories are malleable and subject to distortion.

The same vulnerability appears to be possible for people with highly superior autobiographical memory. In a study led by Lawrence Patihis, the researchers recruited 20 people like A. J. who had demonstrated exceptional autobiographical memory. They were asked to look at slideshows showing a man stealing a wallet or breaking into a car. They then read some descriptions of the slideshows that contained some subtle misinformation. And like Loftus's participants, the people with outstanding memory were just as likely to exhibit memory distortions as control subjects.

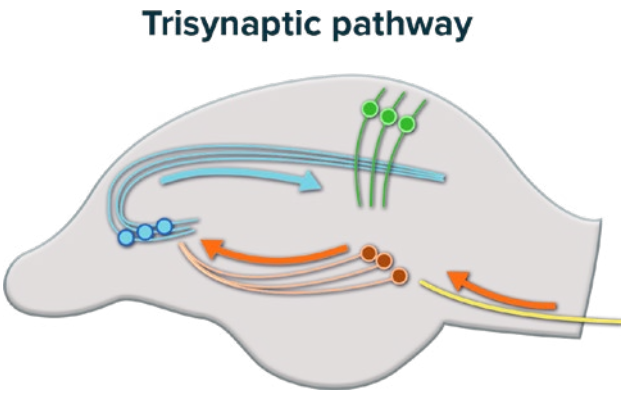
In another experiment, participants were told that the famous crash of United Flight 93 in Pennsylvania had been captured on video, even though it really hadn't. After being fed this misinformation, many normal people and many people with superior memory thought that they remembered seeing the video footage back in 2001. But of course, they hadn't.

Neural Circuits of Episodic Memory

The brain region most strongly associated with episodic memory is the hippocampus. The word *hippocampus* is derived from the Greek words for seahorse. It was given that name because the C-shaped structure of the hippocampus resembles a seahorse. You actually have two hippocampi, one in the left hemisphere and one in the right hemisphere, and they are in the medial part of the temporal lobes.



Neuroscientists have discovered a specific neural circuit—the trisynaptic pathway—that seems to be critical for remembering the specific details of personal experiences. But how can this neural circuit actually implement episodic memory?



Neural networks typically learn by making very small adjustments to hundreds of thousands, if not millions, of synapses. The goal is to try to find a set of synaptic connection weights that lead the network to produce the desired output when given a particular input. Doing so can take a lot of training and time.

But training and time are a problem if you're trying to implement episodic memory because when you're trying to learn and remember an event in your life, you get one exposure—one chance to form a memory. Somehow you need to get a neural network to do one-shot learning.

A second problem exists. Cognitive scientists typically assume that similar inputs should have substantially overlapping neural network representations. That is, many of the same neurons will be active if you're trying to represent similar concepts. Recall that overlapping neural representations allow neural networks to automatically generalize, which is a problem if you're trying to implement episodic memory because you want to keep each of your memories separate. You don't want your memory of last week's trip to the mountains to get muddled up with last month's trip to the beach.

One popular argument among cognitive scientists is that the reason humans have a hippocampus is because they needed an independent memory system that uses different mechanisms to allow it to do one-shot learning and distinguish similar memories.

These problems have led cognitive scientists to argue that episodic memory needs to use completely different mechanisms than those used in other parts of the brain. But what are those mechanisms? One very influential neural network model of episodic memory was inspired by the anatomy of the trisynaptic pathway—the Complementary Learning Systems model. The model uses very sparse representations, not distributed representations across thousands of active neurons. Here, only a small number of neurons are active in response to any given input. As a result, no two events will ever overlap very much, and so different memories can be kept separate from each other without much interference. This process is called pattern separation.

But this model also incorporates a mechanism called pattern completion—when a network will complete a pattern based on only a part of that pattern. So if recurrent connections exist between all the neurons in a layer, and connections are stronger between neurons that are active at the same time, then activated patterns will become attractors for the network. If you give it a part of the pattern, then the network will complete it and retrieve that complete stored pattern.

Such a process is exactly what you want for episodic memory. Given part of an episodic memory, you want to be able to retrieve the rest of that memory. For example, if a particular location, sound, or smell is associated with some event from your past, then you want to be able to retrieve the other details from that event when you visit that location, hear that sound, or smell that scent. And this retrieval is what pattern completion provides.

In this model of episodic memory, the attractor network learns very quickly. The connections among the artificial hippocampal neurons get very strong, very fast. In fact, a single presentation—say, seeing or smelling something just one time—will strengthen the synaptic connections enough to store the attractor pattern, allowing the model to do one-shot learning.

Reading

Hasselmo, M. E. *How We Remember: Brain Mechanisms of Episodic Memory*. Cambridge, MA: MIT Press, 2012.

Hayasaki, E. "How Many of Your Memories Are Fake?" *The Atlantic*, November 18, 2013. https://www.theatlantic.com/health/archive/2013/11/how-many-of-your-memories-are-fake/281558/.

Macmillan, A. "The Downside of Having an Almost Perfect Memory." *Time*, December 8, 2017. https://time.com/5045521/highly-superior-autobiographical-memory-hsam/.

Norman, K. A., G. Detre, and S. M. Polyn. "Computational Models of Episodic Memory." In *The Cambridge Handbook of Computational Psychology*, edited by R. Sun, 189–225. Cambridge: Cambridge University Press, 2008. https://doi.org/10.1017/CBO9780511816772.011.

Wells, G. L., and E. F. Loftus. "Eyewitness Memory for People and Events." In *Handbook of Psychology: Forensic Psychology*, edited by R. K. Otto and I. B. Weiner, 617–629. Hoboken, NJ: John Wiley & Sons, Inc., 2013.

9

Semantic Memory: The Mind's Knowledge Base

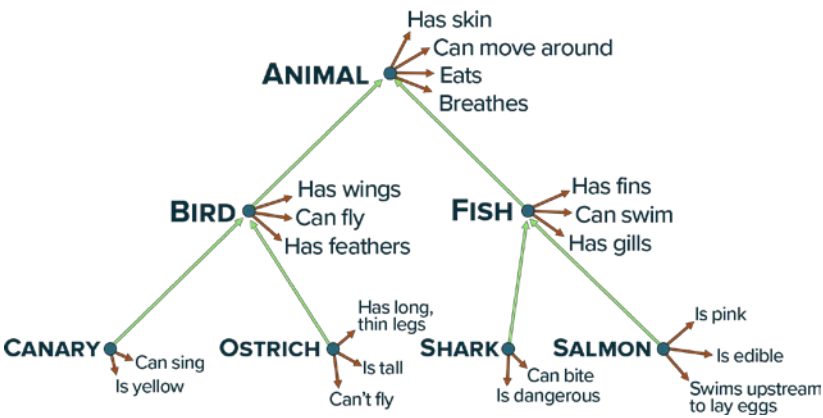
You have storehouses of knowledge about tens of thousands of concepts, from dogs to computers. You didn't come into the world that way, but now you access that enormous database all the time and with incredible accuracy and efficiency. This amazing capability is related to semantic memory—your ability to learn about how the world is organized so that you can recognize, understand, and make inferences about all the things that you encounter in your daily life. In this lecture, you'll learn how cognitive scientists have used various computational models to reveal how semantic memory works.

Hierarchical Semantic Network Models

Recall that episodic memories are tied to a specific time and place and are remembered from a first-person perspective. Semantic memories, on the other hand, are memories for disembodied facts about the world. They correspond to your conceptual knowledge base.

One of the first computationally explicit models of semantic memory was developed by Ross Quillian in Cambridge, Massachusetts. He built a computer program that could answer basic questions about a few simple concepts, such as canaries and salmon. He also teamed up with Allan Collins to test predictions of the model.

The model was based on a hierarchical semantic network. Each node in the network corresponds to a different concept, and each concept has an associated set of features. For example, the concept "canary" is associated with the feature "singing," and the concept "bird" is associated with the feature "flying."



Furthermore, the concepts are organized into a strict hierarchy. For example, "bird" is a higher-level concept—a larger category—and "canary" is a lower-level concept. Lower-level concepts are linked to higher-level concepts via so-called ISA links. The ISA link refers to the connection between the two concepts: If a lower-level concept is a member of a higher-level concept, then it would be connected to that concept via an ISA link.

For example, canaries are a type of bird, and so the "canary" concept is connected to the "bird" concept via an ISA link, indicating that a canary is a bird. Likewise, the "bird" concept is connected to the higher-level concept "animal" because a bird is a type of animal.

And using these ISA links, concepts in this model can inherit properties from concepts that are higher in the hierarchy. For example, if the model knows that birds can fly, then it can infer that canaries can fly based on the ISA link indicating that canaries are birds.

In the hierarchical approach, you can associate a feature with a single concept in the hierarchy, and all the concepts underneath that concept will automatically inherit that feature.

This model provides a relatively simple mechanistic explanation for how humans represent and access semantic knowledge. It also makes some specific predictions that can be tested empirically. A natural assumption would be that it takes time to move up the hierarchy. If so, then inheriting a feature from a higher-level concept should take longer than processing a feature that is directly associated with a concept. Likewise inheriting a feature from a concept that is two levels up in the hierarchy should take longer than inheriting a feature from a concept that is one level up. Early studies confirmed this prediction.

These studies asked participants to decide if sentences were true or false as fast as possible. People were fastest to verify that canaries can sing—a feature associated directly with the concept "canary"—and slowest to verify that canaries breathe—a feature that requires going up two levels in the hierarchy to the level of "animals." And their speed in verifying that canaries can fly—a feature inherited from the concept "bird"—was intermediate between the other two. These results seemed to confirm the basic operation of the model and generated real excitement in the cognitive science community.

However, new experiments exposed serious problems with the original model. Scientists found what are sometimes called reverse distance effects—that is, reaction-time effects that were the reverse of what the original model would predict.

For example, people are faster to verify that a dog is an animal than they are to verify that a dog is a mammal. But dogs are a type of mammal, and mammals are a type of animal, so the concept "mammal" should be between "dog" and "animal" in the semantic hierarchy—meaning that the distance, in terms of links, between "dog" and "mammal" should be shorter than the distance between "dog" and "animal." And so, the model predicts that people should be faster to verify that dogs are mammals because of the shorter distance. But since that did not happen, this reverse distance effect demonstrated that the original model was wrong.

Parallel Distributed Processing Models

One very important advance was in the development of parallel distributed processing models of semantic memory—that is, artificial neural network models in which conceptual knowledge was distributed across populations of neurons that processed the information in parallel. One of the most influential of these models was developed by David Rumelhart at Stanford University.

This model represents the same kind of conceptual knowledge as the earlier hierarchical model, but it actually learns that knowledge via training using the backpropagation algorithm, which was introduced in the discussion about language processing (lecture 5).

The model is a neural network that begins with an item and a relation as inputs and gets matching attributes as outputs. So you start with two input layers. One layer represents an item, such as "canary" or "bird," and one layer represents a relation, such as "can" or "is a." Each input layer contains a bunch of artificial neurons or units, and each one of those units corresponds to a single concept.

Processing in Rumelhart's network is just like the processing you learned about in other neural networks, like ChatGPT and other language networks. First you present specific inputs by setting the activation of one item unit and one relation unit to 1, while setting all the other input units to 0. For example, the "canary" and the "can" units might be set to 1, while all the other input units are set to 0.

Then that activation spreads through some hidden layers and ultimately reaches an output layer that corresponds to different attributes that the item might have. The hope is that the network will activate the correct attribute units and only the correct attribute units. For example, if the item input is "canary" and the relation input is "can," you want the output units corresponding to "grow," "move," "fly," and "sing" to be activated, but you don't want any of the other attribute units to be activated.

But the model has to be trained to do that, which is done by using the backpropagation algorithm. In simple terms, this method is a way to repeatedly make small changes to all the connection weights to get closer to a desired output.

In the parallel distributed processing model, the idea is to match an item and a relation to the corresponding attribute.

Initially, those connection weights are all random, and so the network produces completely random patterns across the attribute layer. But then it compares the pattern it produced to the correct target pattern and computes the error. It then figures out what small change to all the connection weights will reduce the error the most, and it makes those changes.

It repeats this process thousands of times with all the possible input-output pairs. Eventually, the backpropagation algorithm finds a set of connection weights that manages to incorporate all the relevant conceptual knowledge on which it was trained.

Given "canary" and "can" as inputs, the model now produces "grow," "move," "fly," and "sing" as outputs. Likewise, given "rose" and "is a" as inputs, the model produces "flower," "plant," and "living thing" as outputs. For any

given pair of items and relations, the model can then produce the appropriate attributes. In short, it has learned a set of connection weights that reflect all the conceptual knowledge that was hand-coded in the original hierarchical model discussed earlier. But this model does not incorporate a strict hierarchy, and so it doesn't make the same incorrect predictions about distance effects that the Quillian model did.

Similarity-Based Generalization

Rumelhart's model also exhibited other interesting behaviors. In particular, it exhibited similarity-based generalization. That is, it can generalize about the attributes of items it wasn't trained on, based on their similarity to other items that it was trained on. If it was taught something about a new item—for example, that a sparrow is a bird—then the model would make plausible guesses about other characteristics of sparrows, even though the model had never been trained on that information. It would guess that sparrows are living things, that they are animals, and that they can grow, move, and fly.

But how could the model guess so accurately? Think about what happens when you train the model on the fact that a sparrow is a bird. Now you have an item (sparrow) and a relation (is a), and you want to produce an attribute (bird). If you want to activate the "bird" unit in the attribute layer, the model is going to have to learn to produce distributed patterns in the hidden layers that are similar to patterns produced by the other birds in the input because, of course, those are the patterns that excite the "bird" attribute. But if the patterns in the hidden layers are similar to the patterns for other birds, then it will also produce similar activation patterns across the entire attribute layer, including all the nonbird attributes.

For example, other bird patterns tend to activate the "living thing" and "animal" attributes, and so the sparrow pattern will do the same. Likewise, other bird patterns activate the attributes representing that birds can grow, move, and fly, and so the sparrow pattern will also activate those units.

In short, the same hidden units represent all the concepts that activate the "bird" attribute. So all of those bird-activating concepts will tend to have similar hidden unit representations. And because their representations are

so similar, those concepts will also tend to activate other attributes that are typically associated with birds, even if they weren't explicitly trained on those attributes. The model will naturally generalize and make plausible guesses about attributes on which it wasn't trained. And of course, that's exactly what human beings do.

Rumelhart's model also explains why some sets of concepts get categorized together while others don't. When attributes are strongly correlated across different concepts, those concepts tend to be categorized together.

Several models have demonstrated that abstract concepts, such as fairness and justice, can also be learned based on correlations. But rather than arising from correlated attributes, abstract concepts can arise from correlated contexts. If two words, like *fair* and *just*, appear in very similar contexts, then they have to be strongly related. And models that exploit correlated contexts have been able to learn the meanings of abstract words.

Modality-Based Neural Organization

To understand how semantic memory is implemented in the brain, it helps to first consider category-specific semantic impairments. Studies of neurological patients have found that damage to specific parts of the brain can occasionally lead to impairments in knowledge about specific semantic categories.

For example, Elizabeth Warrington and Tim Shallice reported four patients who seemed to have lost their knowledge about living things after brain damage. They had significant trouble providing verbal descriptions of any kind of animate object. For instance, one of the patients described a wasp as a bird that flies and a spider as a person looking for things. These patients also had trouble identifying pictures of living things, such as animals and plants.

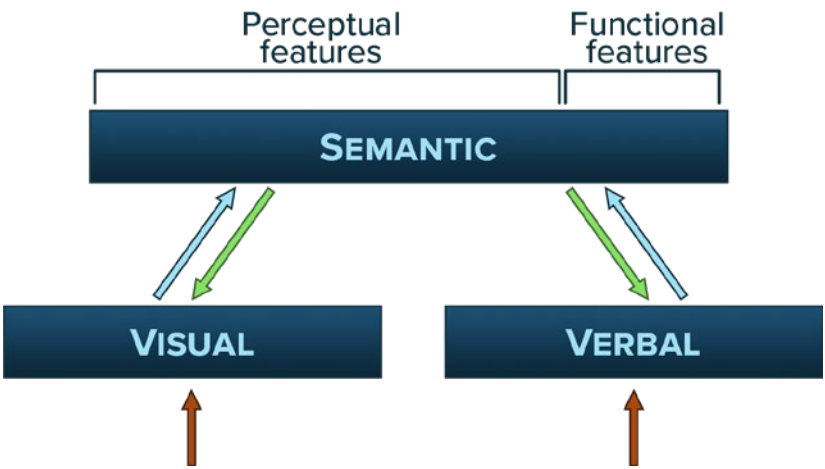
In contrast, these same patients provided quite reasonable descriptions of nonliving things, and they could recognize nonliving things from pictures. Warrington and Shallice pointed out that living things are distinguished mainly by their sensory attributes, while nonliving things are mainly distinguished by their functional attributes. For example, think about what you know about an elephant. It's big and gray, and it has floppy ears and a long trunk. Those are all sensory features—that is, what it looks like.

But now think about what you know about a nonliving thing, such as a car. In addition to visual features, you also know that you need to start the car, put it in gear, and use the pedals to accelerate and brake so that you can get from point A to point B.

A lot more functional information is associated with nonliving things than with living things. So maybe semantic memory isn't organized by category but rather by modality. Maybe sensory information is stored in one part of the brain, while functional information about how you might interact with the object is stored somewhere else. According to that theory, if the brain regions storing sensory information are damaged but those storing functional information are not, then you might have a hard time conceptualizing living things.

Martha Farah demonstrated the plausibility of this hypothesis by instantiating it in a neural network model with her colleague Jay McClelland. The model has three layers. Think of them as three groups of artificial neurons that represent three different categories of information: visual, verbal, and semantic.

Patterns of activation across the visual layer represent different pictures, while patterns across the verbal layer represent different verbal labels. Patterns in the semantic layer represent different concepts, some of which are living things and some of which are nonliving things.



Farah and McClelland assigned all these patterns randomly—with one important exception. The neuronal units in the semantic layer were further subdivided between two kinds of information: perceptual and functional. They trained the model so that it could produce the correct semantic pattern when given either a visual pattern or a verbal pattern. When training was done, the model could generate the correct semantic pattern for any input.

When perceptual semantic units were damaged, the model exhibited what looked like category-specific semantic impairments for living things. Conversely, when functional semantic units were damaged, it displayed what looked like a selective impairment for nonliving things.

These results demonstrated that Warrington and Shallice's idea could work in practice. Semantic impairments may be category-specific, but this doesn't imply that semantic memory is organized by category in different parts of the brain. A more plausible model is a neural network in which semantic memory is organized by modality. That is, some regions of the brain represent perceptual features, and other regions represent functional features.

Reading

Hart, J., and M. Kraut, eds. *Neural Basis of Semantic Memory*. Cambridge: Cambridge University Press, 2007. https://doi.org/10.1017/CBO9780511544965.

McRae, K., and M. N. Jones. "Semantic Memory." In *The Oxford Handbook of Cognitive Psychology*, edited by Daniel Reisberg, 206–220. New York: Oxford University Press, 2013. https://doi.org/10.1093/oxfordhb/9780195376746.013.0014.

Rogers, T. T. "Computational Models of Semantic Memory." In *The Cambridge Handbook of Computational Psychology*, edited by R. Sun, 226–266. Cambridge: Cambridge University Press, 2008. https://doi.org/10.1017/CBO9780511816772.012.

Rogers, T. T., and J. L. McClelland. *Semantic Cognition: A Parallel Distributed Processing Approach*. Cambridge, MA: MIT Press, 2004.

10

The Animal Mind

This lecture focuses on two aspects of mental life that have sometimes been claimed to distinguish human beings from other animals: language and consciousness. You'll learn about fascinating work that has been done investigating animal cognition, including studies that aimed to answer the following questions: Are nonhuman animals capable of learning a language, and do they have conscious experience like humans do?

The Clever Hans Effect

On September 4, 1904, *The New York Times* published an article entitled "Berlin's Wonderful Horse." The author provided an eyewitness account of the amazing performance of a black Orlov trotter stallion named Hans and his owner Wilhelm von Osten.

Herr von Osten had spent 4 years training Hans for these public performances. Afterward, the horse seemed to be able to perform some truly remarkable cognitive tasks. For example, when von Osten put an arithmetic question to Hans, the horse would tap out the answer with his hooves (once for one, twice for two, and so on). Every time, Hans got the answer right! He could even point to letters and spell out words.

The article goes on to describe other amazing intellectual feats, including indicating the color of different objects, providing the correct time by reading a watch, and providing the correct month and day of the week. Hans became a media sensation, and tourists constantly stopped by to see the famous horse, who was now typically referred to as Clever Hans.

The scientific community, however, was suspicious and wanted to test Clever Hans more thoroughly. So the philosopher and psychologist Carl Stumpf organized a commission of 13 experts, including a veterinarian, a circus manager, and the director of the Berlin Zoo. They tested Hans over a period of 2 days, including numerous tests when his owner wasn't even there.

And Hans passed the tests! The commission saw Hans consistently answer their questions correctly, even when von Osten wasn't around, and they unanimously concluded that Hans was performing these amazing feats without any kind of assistance.

So what was really going on? The truth finally came to light a few weeks later when Oskar Pfungst, an assistant in Dr. Stumpf's lab, discovered that although Hans could answer questions without von Osten being around, he could not do so when he had blinders on that prevented him from seeing the person asking the question. Likewise, Hans's performance was terrible when the person asking the question did not know the answer.

Pfungst concluded that all the people questioning Hans made very subtle changes in their posture and/or facial expression when the horse made the final, correct stamp with his foot. And apparently, Hans was able to detect these unconscious changes and use them as a cue for when to stop tapping. With blinders on, Hans never saw the nonverbal cues, and his miraculous abilities suddenly disappeared.

These findings are now often referred to as the Clever Hans effect, referring to the fact that animal behavior can sometimes be influenced by very subtle cues in the behavior of the people performing the tests. And these results have had a big influence on how animal cognition is studied. Researchers now try their best to test animals in ways that can't be influenced by human cues.

Studying Language in Animals

All nonhuman animals communicate, and many do so in very sophisticated ways. Bees perform waggle dances to communicate the direction to a food source as well its distance. Monkeys yell out specific cries to warn other monkeys about nearby predators, and these cries can be different depending on the predator.

But do examples like these constitute language? Most cognitive scientists would say no. Real, natural languages involve producing and comprehending combinations of words, even if those combinations of words have never been produced or comprehended before. So while there's agreement that nonhuman animals can and do communicate in many different ways, it's an open question as to whether they can learn to understand and produce what cognitive scientists consider to be language.

A number of cognitive scientists have tried to answer that question by training animals to respond to and produce language. One of the first such efforts was conducted by Keith and Catherine Hayes. They tried raising a chimpanzee named Viki as much like a human child as possible. They wanted to immerse her in a traditional human environment and see if it might lead Viki to develop any language. Viki lived in the Hayes's home for 3 years. She wore

diapers, ate at the table, and was exposed to human language throughout her upbringing. The Hayeses also tried their best to teach Viki to say human words.

Unfortunately, Viki didn't learn much language. In fact, she only ever managed to produce four sounds that the Hayeses interpreted as the words *mama*, *papa*, *up*, and *cup*. So what went wrong? One possibility is that Viki didn't have the cognitive capacity to learn much human language. Another possibility is that she didn't have the vocal capacity. The chimpanzee larynx is quite different from the human larynx, and studies suggest that it may not be capable of producing the sounds required for human speech.

But even though chimpanzees and other great apes don't have the necessary vocal machinery to produce human speech, they do have the manual dexterity to produce many human signs. So several researchers began trying to teach American Sign Language to nonhuman animals.

Allen and Beatrix Gardner at the University of Nevada, Reno, were among the first. They raised a chimpanzee named Washoe in their home and tried to mimic the way a human child would be raised while simultaneously teaching her American Sign Language. They reported that she learned more than 100 signs.

Probably the most famous case of teaching a great ape language involved Penny Patterson and her gorilla Hanabiko, or Koko, as she was normally called. Like the Gardners, Patterson also reported that after a little more than a year of training, Koko had learned more than 100 signs from American Sign Language. Over the next few years, that estimate increased to thousands of signs with additional training. And Koko seemed to combine signs in novel ways. For example, when exposed to a magnet, she produced the signs "stuck" and "metal," and when presented with a ring, she signed "finger" and "bracelet."

Koko became a media sensation. Video clips demonstrated a real emotional bond between Koko and Patterson and appeared to show humanlike conversations in sign language with Patterson translating the signs that Koko produced. Such videos are enough to convince many people that great apes

can indeed be taught to use sign language to communicate effectively with human beings. Unfortunately, there isn't much solid scientific evidence to back up that claim.

For one thing, most of the people who trained these animals have failed to make the raw data—including raw video footage—available to other researchers. Furthermore, other researchers who have tried to train great apes to use sign language have reported much less impressive results. The case of Nim Chimpsky, named after the famous linguist Noam Chomsky, was particularly influential. Nim was a chimpanzee taught by Herbert Terrace and his colleagues. They reported that Nim learned more than 100 signs. But, of course, real language use involves more than knowing individual words—it involves combining words together in grammatical ways.

So Terrace and his colleagues analyzed more than 19,000 multisign sequences that Nim produced to see whether the signs had some systematic order or were just random sequences with no obvious design. If the order was systematic, it would indicate that Nim really could understand and produce sign language.

The results looked promising. There was clear systematicity in the order of the signs that Nim produced. Just like human English speakers say "more food" rather than "food more," Nim was much more likely to use the sign "more" before a sign for some kind of food than to reverse the words. Based on the systematicity of Nim's two-word utterances, you could argue that he had learned some rudimentary grammar rules.

However, Nim's three-word utterances looked less human and typically included unnecessary redundancy. He would say things like "play me Nim" and "eat me Nim" instead of just "play me" or "play Nim." And longer utterances did not typically add new information or elaborate on an idea like they would in real human language.

A second problem was that the individual signs produced by the animals weren't typically accurate by real sign language standards. Human beings had to make an educated guess about what the signs might actually be. And people who knew American Sign Language well sometimes thought the human interpreters were far too generous.

A third problem that the researchers reported was that most of Nim's signs were not spontaneously generated; rather, they were imitations of signs that the trainer had just used. In normal human language, speakers take turns and wait for the other speaker to finish talking before responding. Not so with Nim and other signing apes. In contrast, human children learning American Sign Language almost never start signing before the other person is finished. And their responses typically go beyond what they just observed and expand on the ideas in some way.

Most cognitive scientists today think what was really going on with Nim and the other apes was just a Clever Hans effect. After years of training, they had learned that making certain hand movements at certain times led to rewards. And since repeating the hand movements of their trainers often seemed to work, that's what the apes learned to do. It's very unlikely that they had learned how to combine words together in novel ways to convey ideas, even if that's what their trainers thought they were doing. It's more likely that their trainers couldn't help but anthropomorphize the apes and attribute humanlike language abilities to them, just like Clever Hans's trainer had done more than 50 years earlier.

This argument raises another important question. If human beings tend to attribute humanlike language abilities to animals that don't actually possess them, do they also do that with other aspects of human experience, such as consciousness?

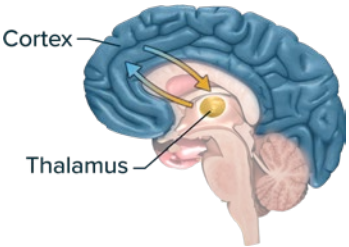
Studying Consciousness in Animals

Most pet owners would adamantly argue that their pets exhibit evidence of conscious experience. But is there any scientific evidence to support this kind of attribution? It turns out that there is some relevant evidence.

However, the topic of animal consciousness is still quite contentious, and no clear consensus exists on which nonhuman animals, if any, possess consciousness and whether that consciousness is anything like the subjective experience that humans possess. In particular, the scientists and philosophers who specialize in consciousness have a hard time even defining the concept

of consciousness in a way that everyone agrees on. So as you consider the evidence presented here, recognize that many cognitive scientists are not yet convinced.

One type of evidence comes from neuroscience. Anil Seth, Bernard Baars, and David Edelman at the Neurosciences Institute in San Diego reviewed evidence that consciousness in humans is associated with low-amplitude interactions in the circuits connecting the thalamus and cortex. These interactions are both fast and widespread. They also happen to be present in most mammals, suggesting that these nonhuman species probably also have some kind of conscious experience.



Probably the most cited evidence for animal consciousness comes from a paradigm that was developed by Gordon Gallup at Tulane University in the late 1960s. Gallup examined the behavior of four chimpanzees when a mirror was placed outside their cage so that they could see themselves. Initially, the animals seemed to respond as if the mirror image was another chimpanzee. But over time, the social responses were replaced by self-directed behaviors that used the mirror, such as grooming themselves, cleaning their teeth, and picking material out of their nose.

Intrigued by this behavior, Gallup conducted an experiment. He anesthetized the animals and marked one of their ears and one of their eyebrow ridges with red dye. He let the dye dry so that the chimps would not be able to feel it or smell it when they woke up. Then, he removed the mirror and returned the animals to their cage. He watched the chimpanzees for 30 minutes to see if they touched the spots he had marked with the red dye. In general, in the absence of the mirror, they seemed oblivious to the red dye.

But as soon as Gallup put the mirror back, the chimps' behavior changed dramatically. They spent more time looking at themselves in the mirror than they did during the period before the dye had been applied. They touched the

spots with the dye more often than they had when they didn't have a mirror. They also sometimes touched the dye and then looked at or smelled their fingers, even though the dye was dry and didn't transfer in any way.

This study provides compelling evidence that chimpanzees have some level of self-awareness, a critical component of consciousness. And it turns out that chimpanzees are not alone. Other great apes, such as orangutans and bonobos, have also passed Gallup's mirror test. And so have dolphins, elephants, and even magpies. On the other hand, many other species, such as dogs, pandas, and many types of monkeys, have failed the test. Do those animals not have self-awareness?

Although some cognitive scientists have tried to make that argument, most would not. The mirror test obviously depends crucially on vision, and species that put more emphasis on other senses, such as smell, might fail the test even if they are self-aware. Still other species have a strong instinct to avoid eye contact with other members of their species and may therefore avoid looking at the mirror. And other species may not care about the mark even if they are aware that it is on them. The bottom line is that although passing the mirror test provides evidence for some kind of self-awareness, failing to pass the test doesn't necessarily imply a lack of self-awareness.

Reading

Andrews, K. *The Animal Mind: An Introduction to the Philosophy of Animal Cognition*. London: Taylor & Francis, 2014.

de Waal, F. *Are We Smart Enough to Know How Smart Animals Are?* New York: W. W. Norton & Company, 2016.

Griffin, D. R. *Animal Minds: Beyond Cognition to Consciousness*. Chicago: University of Chicago Press, 2013.

Morell, V. *Animal Wise: How We Know Animals Think and Feel*. New York: Crown, 2014.

11

The Psychology of Decision-Making

Suppose you're about to flip a fair coin four times. Do you think that you're more likely to flip a head followed by two tails followed by another head or that you're more likely to flip four straight heads? Most people assume that the first sequence is more likely than the second. But both outcomes are equally likely. This scenario is just one example of how fallible a human being's judgments can be. This lecture explores how intelligent, rational creatures regularly make what seem like irrational decisions. In particular, it examines some descriptive theories that try to explain people's mistakes in terms of bounded rationality and shortcut heuristics.

Conditional Probability

Mammogram results come back positive in about 80% of patients who have breast cancer. But that does not mean that a patient who has a positive mammogram has an 80% chance of having breast cancer. To understand why, you need to first understand conditional probabilities.

Probability is the likelihood that some event will occur or that some proposition is true. A conditional probability is the likelihood that some event will occur assuming that some other event has already occurred.

For example, suppose you give mammograms to a group of women who have already been diagnosed with breast cancer. What is the probability that the mammograms will come back positive given that they do have breast cancer? Such a scenario is a conditional probability. And that probability is around 80%. So, 80% of patients with breast cancer will test positive on a mammogram.

But here's the critical point: That conditional probability is not the one that the doctor in the example is trying to figure out. The doctor and the patient want to know the probability of breast cancer given that the patient has a positive mammogram. In fact, the probability is only about 7%. Unfortunately, many people, including many doctors, assume these two conditional probabilities are the same, and they obviously aren't.

Why do people tend to confuse the two? To understand this kind of faulty reasoning, you first need to understand the difference between normative theories and descriptive theories.

A normative theory explains how people should behave if they're performing the task in the best possible way. In mathematics, a normative theory would explain how to arrive at the correct answer to a problem. In a probabilistic reasoning task like medical diagnosis, a normative theory tells you what information the doctor should consider and how they should incorporate that information to arrive at the true probability of a diagnosis.

Normative theories describe how people should behave, while descriptive theories describe how people actually behave.

But people usually aren't normative. They get confused about the information they should consider all the time. Ultimately, cognitive scientists want to understand how people actually behave, not just how they should behave. And descriptive theories try to explain actual human behavior, including all the stupid mistakes and errors that people are prone to make.

Expected Utility Theory

One of the most influential normative theories is called expected utility theory. According to this theory, any particular decision might have several potential outcomes. So suppose you have a numeric measure of how good each outcome is, which can be referred to as its utility. And suppose that you know how likely each potential outcome is, that is, it's probability. You can then multiply the probability of each outcome by its utility and then add up all of those products, which will give you the so-called expected utility of that choice.

So, for instance, if there's a high probability of a good outcome given some choice, you could say that choice has a high expected utility. And if you do the same thing with all the other choices, then you can pick the choice that has the highest expected utility.

Decision	Outcomes	Utility	Expected Utility
A	100% chance you win $1	$1	$1
B	50% chance you win nothing and 50% chance you win $3	$0 $3	$1.50

For example, suppose that you have to decide between two alternatives, A and B. If you choose A, then you will be guaranteed to win $1. If you choose B, then there's a 50% chance that you will win nothing and a 50% chance that you will win $3. Which would you choose?

In this case, most people choose B over A. Even though the probability of winning money is only 50% in option B, the utility of $3 is much higher than the utility of $1. In particular, if you assume that the utility corresponds to the monetary value, then the expected utility of option A is 100% multiplied by $1, which is $1, while the expected utility of option B is 50% multiplied by $3, which is $1.50. So in this case, option B has a higher expected utility, and it makes sense that people would choose it.

But it's important to point out that utility is a personal choice and doesn't necessarily map onto monetary value. For example, suppose you change the problem such that if you choose option A, you are guaranteed to win $1 million. And if you choose option B, there's a 50% chance you'll win $3 million, but there's also a 50% chance you'll win nothing.

In that case, almost everyone chooses option A and goes for the guaranteed $1 million, even though the expected monetary value of option B is higher; it's $1.5 million. So why do people prefer option A?

The decision can be perfectly consistent with expected utility. One natural explanation is that the utility of $1 million is already very high, and the utility of $3 million is not three times as large. So it's not worth that 50% risk to try to go from $1 million to $3 million.

For example, you might say that winning $1 million has a utility of 10 on your own personal, arbitrary scale. On that same scale, the utility of $3 million might be 13 or 14, but it's not going to be much higher than that. The expected utility of option A is 100% multiplied by 10, which is 10. And the expected utility of option B is 50% multiplied by 13 or 14, which is around 6 or 7. And so, expected utility theory can explain why it is appropriate for you to choose option A in this case, even though you chose option B when you were deciding between $1 and $3.

Calculating Conditional Probabilities

There is a normative approach to figuring out conditional probabilities like the probability of breast cancer given that a mammogram came back positive. The approach is called Bayes's theorem, after the 18th-century Presbyterian minister who first discovered it. In mathematical form, Bayes's theorem corresponds to the following equation:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

This equation has four probabilities, represented by *P*s, followed by something in parentheses. The left side of the equation stands for the conditional probability of *A* given *B*. Using the earlier example, you can say *A* is the actual presence of breast cancer, and *B* is a positive mammogram. So the left side of the equation should tell you the probability of breast cancer, assuming you have a positive mammogram result.

The right side of the equation has three probabilities, two in the numerator on the top and one in the denominator on the bottom. The first one on the top is the probability of *B* given *A*—that is, the conditional probability of *B* assuming that *A* is true. In the mammogram example, this might be the 80% sensitivity of the mammogram result: Assuming breast cancer is present, the probability of getting a positive mammogram is around 80%.

The other probability in the numerator is just the probability of *A*. This probability is not a conditional probability but is the base-rate probability of *A* independent of any other events—also often called the prior probability, meaning the probability before considering any other events. In the example, *P*(*A*) would correspond to the base-rate probability of breast cancer—basically, how common the disease is.

Finally, the probability in the denominator is another base-rate probability, namely, the probability of *B*—in the example, the probability of getting a positive mammogram in the population as a whole.

Remember, Bayes's theorem is normative. It tells you the correct way to calculate a conditional probability, like the probability of breast cancer given a positive mammogram. And so, Bayes's theorem tells you exactly the information that you need to consider when trying to decide on a conditional probability like a diagnosis.

And what it tells you is that you need to consider both how likely the symptom is given the disease and the base-rate probability of the disease itself. In the breast cancer example, the doctor needs to consider not only that mammograms will come back positive in 80% of patients with breast cancer but also how common breast cancer is.

Descriptive Theories

One of the most important ideas behind descriptive theories in psychology was proposed by Dr. Herb Simon at Carnegie Mellon University in the 1950s. He argued that although human beings are rational, their rationality is bounded in some important ways.

For one thing, their knowledge is bounded. For example, suppose you want to buy a winter coat, and three stores nearby sell coats. Ideally, you'd go to the store that has a coat you really like at the cheapest price. Expected utility theory would predict you'd do this every time. The problem, of course, is that you don't know beforehand which store has that coat.

So you'd probably end up going to one of the stores and looking through their coats. And if you found one that you liked and thought was priced appropriately, then you'd buy it. Of course, the same coat might be on sale for less at another store, but you don't know that. If you did, then your behavior seems kind of irrational. But given the bounds of your knowledge, it seems perfectly reasonable.

Simon referred to this kind of behavior as *satisficing*, which is a blend of the words *satisfy* and *suffice*. You regularly make choices that would not be considered optimal if you had complete knowledge of the situation, but given the bounds of your knowledge, your choices satisfy your goals and are good enough even if they're not strictly optimal. People also have bounded time

and bounded cognitive abilities. In the coat example, you probably don't have time to research every coat at every store, and even if you did, you may not remember all the information that could be relevant to your decision.

How Heuristics and Framing Affect Judgment

Bounded rationality also applies when you're trying to figure out probabilities. For example, suppose you're eating a meal with someone, and you're trying to guess how likely they are to have the same opinion as you on some controversial topic involving politics or religion. You don't want to ask them directly because it might be awkward, but you'd like to know because it could affect future interactions in some way.

Many people unconsciously apply what is sometimes called the representativeness heuristic. A heuristic is a strategy that tends to produce a reasonable answer quickly even though it's not guaranteed to produce the optimal answer. And people often use the representativeness heuristic to estimate probability based on similarity or representativeness. So, for example, if the person you're eating with seems representative of people who hold a particular opinion on the controversial topic in question, then you might assume that they probably hold that opinion, too.

Representativeness could also explain the thinking in the positive mammogram example. Patients with breast cancer tend to have positive mammograms. So if a new patient has a positive mammogram result, that patient seems similar to patients with breast cancer. And since the new patient seems representative of patients with breast cancer, someone might assume that there must be a high probability that the new patient has breast cancer, too.

One of the main reasons why the representativeness heuristic leads people astray is that it completely ignores base-rate probabilities. Remember from Bayes's theorem that if you want to figure out the conditional probability of *A* given *B*, you need to consider both the conditional probability of *B* given *A* and the base-rate probabilities of *A* and *B*.

So if you want to estimate the probability of breast cancer given a positive mammogram, you can't just think about the probability of a positive mammogram given breast cancer. You also have to consider the base-rate probability of breast cancer to begin with. And that base-rate probability is actually quite low. As a result, even if a patient has a positive mammogram, the probability that they have breast cancer is only about 7%.

Another common heuristic that people use to make probability judgments is called the availability heuristic. When people use this strategy, they're estimating frequency based on how easy it is for them to bring examples to mind from memory. Frequent events should be easier to bring to mind than infrequent events, and they usually are. However, factors other than frequency can also influence how easy it is to bring examples to mind.

Availability could also explain why people tend to think that breast cancer is more common than, say, diabetes. Breast cancer has received more attention in the popular press than diabetes. Think about pink-ribbon campaigns or NFL players wearing pink cleats to raise breast cancer awareness. These kinds of efforts make breast cancer particularly available and easy to bring to mind. And availability can bias people's judgments of frequency.

Another very common source of bias is the way a problem is framed. Generally, people tend to be risk averse when problems are framed positively, and they tend to be risk seeking when problems are framed negatively. People also typically report that the pain of losing $50 over a coin toss is stronger than the joy of winning the same amount, which is typically referred to as loss aversion in the decision-making literature. Loss aversion leads people to favor choices that avoid losses, even if the riskier choices might have a higher expected value than the alternatives.

Clearly, when humans make decisions, they're not always perfectly rational—they don't always consider only the information that they should consider from a normative perspective—which is why descriptive theories in cognitive science are useful. They give scientists an insight into how people really think and behave when faced with problems that aren't always straightforward. The fact is, human rationality is bounded, and people often rely on heuristic strategies that are fast and typically produce reasonable answers but that also can lead to systematic biases.

Reading

Kahneman, D. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux, 2011.

Hastie, R., and R. M. Dawes. *Rational Choice in an Uncertain World: The Psychology of Judgment and Decision Making*. Los Angeles: SAGE Publications, 2010.

Schwartz, S. "Heuristics and Biases in Medical Judgment and Decision Making." In *Applications of Heuristics and Biases to Social Issues*, edited by L. Heath, R. S. Tindale, J. Edwards, E. J. Posavac, F. B. Bryant, E. Henderson-King, Y. Suarez-Balcazar, and J. Myers, 45–72. Boston, MA: Springer, 1994.

12

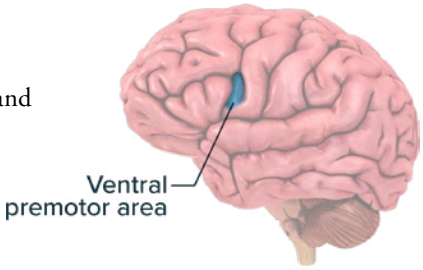
Decision-Making at the Neural Level

This lecture continues the discussion of decision-making by exploring what's happening in the brain at the neural level as decisions are being made. Although many questions still remain about the exact mechanisms, neuroscientists have made substantial progress in understanding the neural basis of simple perceptual decision-making. You'll learn about those studies as well as others that have investigated processes that might help people make better choices.

Neural Activity during Vibrotactile Decisions

Scientists Ranulfo Romo, Adrián Hernández, and Antonio Zainos conducted an experiment at the National Autonomous University of Mexico in Mexico City in which they trained monkeys to perform a simple vibrotactile discrimination task. The animals received two successive vibrations on one of their fingers and then had to indicate which vibration had a higher frequency.

The frequencies and the order of the vibrations varied. Sometimes the vibration frequencies were high, and sometimes they were low. Sometimes the faster frequency came first, and sometimes it came second.



Once the monkeys got good at the task, the scientists then implanted electrodes and started recording the activity of neurons in an area of the brain called the ventral premotor area. Previous studies suggested that this area might be involved in making vibrotactile decisions like this.

When Romo, Hernández, and Zainos looked at the recordings, they discovered that different neurons seemed to be involved in different stages of the decision process. They created a raster plot, which is a graphical display of the activity of a single neuron.

Time runs from left to right, and each of the dots indicates that the neuron fired at that point in time. In addition, each row of dots corresponds to a single trial in which two vibrations were presented. This figure contains a total of 50 rows, so this is plotting the neuron's activity on 50 different trials.

Each set of five rows of dots corresponds to a different pair of frequencies. The first five rows are labeled 26:34. Those numbers refer to frequency. In these trials, the frequency of the first vibration was 26 stimulations per second, while the frequency of the second vibration was 34 stimulations per second. In the top five sets, the frequency of the second vibration, which is labeled f2, is faster than the frequency of the first vibration, which is labeled f1. Conversely, in the bottom five sets, f1 is faster than f2.

Finally, the colored vertical bands indicate the different time periods in every trial. For example, the vertical white band of dots on the left corresponds to the time before the first vibration was delivered. The vertical brown band next to it corresponds to the time when the first vibration was presented. The large blue band in the middle corresponds to the delay period between the two vibrations. The vertical purple band on the right corresponds to the presentation of the second vibration. And finally, the white vertical band on the far right corresponds to the end of the trial, when a monkey is making its response.

The *KU* in the top right stands for "key up" and indicates the point in time when the monkey lifted its finger off a key to begin making a response. Likewise, *PB* stands for "push button," which is the time when the monkey pushed a button to indicate its response.

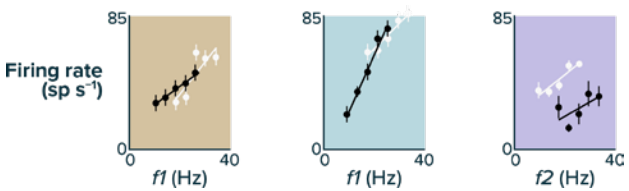
The raster plot revealed that one neuron was firing a lot when the monkey's finger was being vibrated. The neuron also seemed to be encoding information about the frequency of vibration—the highest frequencies produced more activity than the lowest frequencies.

Information about the frequency is obviously the first thing you would need to know if you're going to decide which vibration had a faster frequency. The neuron's activity reflected the vibration frequency only when the stimulation was actually happening. The activity didn't seem to reflect any kind of memory about the first vibration; it just represented a piece of important perceptual information as it happened.

Interestingly, a raster plot from a different neuron in the same ventral premotor area revealed very different behavior, especially during the delay period between the two vibrations and during the period when the second vibration was presented. Unlike the first neuron, the second neuron's activity still reflected the frequency of the first vibration—the firing rate was higher if that vibration was fast and lower if that vibration was slow.

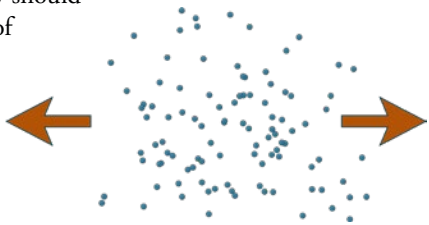
So this neuron was maintaining information about the first vibration's frequency across a delay, just like a working memory. And obviously, you'd need to be able to remember the first vibration's frequency if you were going to compare it to the second vibration's frequency.

So the second neuron seemed to reflect an actual judgment or decision. It fired significantly more at the point when a subject should indicate that the first frequency was higher than the second. The neuron seemed to remember the first vibration's frequency during a short delay and then fired more or less depending on which vibration had a higher frequency, which is exactly the decision that the monkey is trying to make.



Neural Activity during Motion-Based Tasks

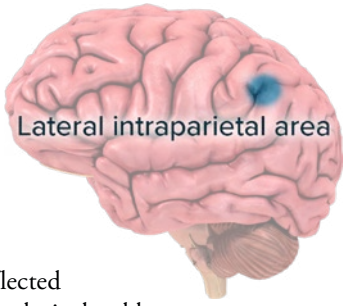
Jamie Roitman and Michael Shadlen at the University of Washington conducted an experiment in which they trained monkeys on a random-dot-motion discrimination task. The task was to decide which direction most of the dots were moving. If most of the dots were moving to the right, then the monkey should have responded "right." And if most of the dots were moving to the left, then the monkey should have responded "left."



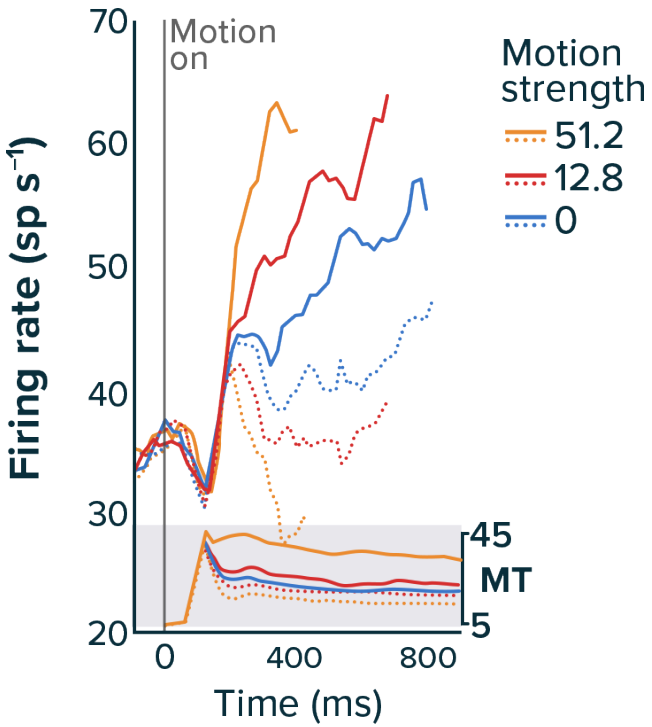
The percentage of dots that are moving in the same direction is referred to as the coherence or strength of the dot motion. If almost all the dots are moving in the same direction, then coherence and strength are high, and the task is easy. Conversely, if a small percentage of the dots are moving together, then the coherence and strength are low, and the task is hard.

By manipulating the motion coherence, scientists can easily change how hard the task is and how long it takes the monkey to respond. And if they record from neurons during the period when the monkey is trying to make a decision, they can watch to see how neural activity changes while the monkey is making the decision.

Roitman and Shadlen recorded from a region of the brain known as the lateral intraparietal area (LIP) as the monkey tried to decide whether most dots were moving left or right. They found that the neurons they examined fired a lot more when the monkey thought the dots were moving to the right, which was the correct response, than when they thought the dots were moving to the left. This activity reflected more about the decision that the monkey thought it should make than about the actual motion of the dots.

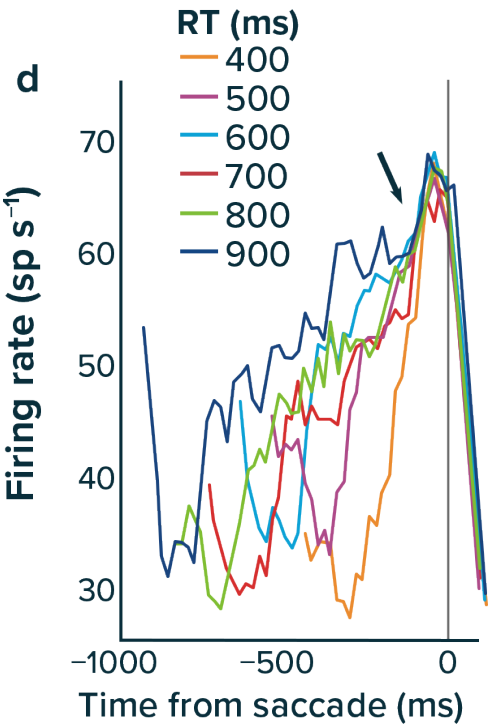


The activity in the neurons also ramped up over time. And the rate at which the activity ramped up reflected the strength of the motion—when the motion was strongly coherent, the neural activity ramped up very quickly. Many neuroscientists hypothesize that this ramping up corresponds to evidence accumulation. The idea is that this neural activity reflects the monkey's belief that there is rightward motion in the dots. And as the monkey gains additional evidence in support of this hypothesis, its belief grows stronger, and so does the neural activity.



So what determines when the monkey will stop accumulating evidence and make a choice? When scientists plotted the activity of the same LIP neurons relative to the time of the decision rather than to the time the dots started moving, they observed two things. First, they noticed that, once again, the ramping-up activity occurred over time. For short-reaction-time trials, the activity ramped up very fast, presumably because there was strong motion coherence, and so the evidence for rightward motion was very strong. Conversely, for slow-reaction-time trials, the activity ramped up more slowly, presumably because the motion coherence was weaker in most of those trials.

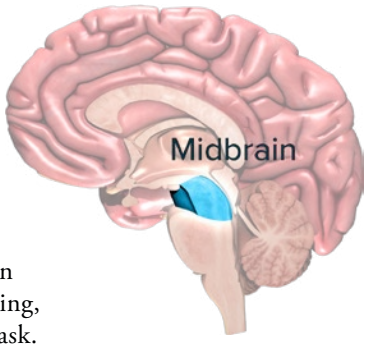
Importantly, the second thing they noticed was that the activity level was always about the same—in this case, around 65 neural firings, or spikes, per second—at the 0 time point, when the monkey actually made the decision and responded. Regardless of how fast or slow the trial was, once the firing rate reached the threshold, the monkey made a response.



Many neuroscientists interpret these findings as evidence for a decision threshold. The idea is that the animal continuously accumulates evidence in favor of a particular decision. And that evidence passing some particular threshold is what leads the animal to commit and make a response.

Neural Activity in Value-Based Decisions

Neuroscientists have also conducted studies to identify neurons that encode value and that might play a role in decision-making. One famous study involved the midbrain region and examined the behavior of a special set of neurons that use the neurotransmitter dopamine. It was conducted by Dr. Wolfram Schultz, who was at the University of Fribourg in Switzerland at the time. In the study, monkeys were rewarded with a squirt of delicious fruit juice if they pressed a lever immediately after a light turned on. Midbrain dopamine neurons were recorded before, during, and after a monkey learned to perform this task.



When the monkey got juice before it learned the task, Schultz observed that the dopamine neuron fired a lot immediately after the monkey got the juice. After the monkey learned the task, Schultz observed that the dopamine neuron started firing rapidly earlier—immediately after the light appeared. The light had become a conditioned stimulus, which is a stimulus that has become strongly associated with a reward by being repeatedly paired with it. In this case, the conditioned stimulus was the appearance of the light that had been repeatedly paired with the juice reward. Just like Pavlov's dogs were conditioned to expect food when they heard a bell, the monkey had been conditioned to expect the juice whenever the light appeared.

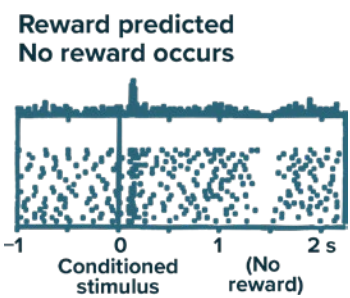
Schultz found that once the monkey had learned that the light predicts the juice reward, the dopamine neuron behaved very differently than it had at first. The neuron now responded to the light rather than to the juice. That is, Schultz saw a big burst of neural activity right after the conditioned stimulus, but he didn't see a similar burst right after the reward.

One popular hypothesis for why that occurred is that these dopamine neurons signal what is sometimes called reward prediction error. That is, the dopamine neuron fired when the monkey received an unexpected reward or when the monkey perceived an unexpected stimulus that was strongly associated with reward.

In this case, the monkey knew that it was about to get a reward as soon as the light turned on. But before the light turned on, the monkey wasn't thinking it was about to get a reward. So when the light came on unexpectedly, the monkey needed to update its reward prediction. In other words, its prediction that it would not be receiving a reward was an error, and it needed to update that prediction. And the dopamine neuron signaled that reward prediction error. It fired rapidly to indicate that an unexpected reward was on its way.

The neuron did not produce a burst of activity when the juice reward was delivered because the monkey was now expecting the reward. After lots of training, the monkey knew that pressing the lever after the light turned on was essentially guaranteed to lead to juice, and so it was expecting the juice reward. And since that prediction was correct, there was no reward prediction error, and therefore the dopamine neuron didn't fire.

Something else happened when the light came on but the monkey didn't get the juice reward. Once again, the dopamine neuron fired a lot right after the conditioned stimulus, which was the light. Everything seemed normal until the moment in time when the monkey would usually receive the juice reward. It was expecting juice, but there was no juice, and suddenly, the neuron's activity dropped off. Such an effect makes sense if you think about the activity as a reflection of reward prediction error.



A reward prediction error signal like this can be used to help you make better decisions that maximize utility. In particular, you can use reward prediction errors to continuously update your estimates of the utility or value of different choice options, which allows you to learn to make better choices as you gain more experience. This idea has been used in the field of artificial intelligence to build systems that exceed human-level capability based purely on learning from experience.

Reading

Glimcher, P. W., and E. Fehr, eds. *Neuroeconomics: Decision Making and the Brain*. Boston: Academic Press, 2013.

Gold, J. I., and M. N. Shadlen. "The Neural Basis of Decision Making." *Annual Review of Neuroscience* 30, no. 1 (2007): 535–574.

Redish, A. D. *The Mind within the Brain: How We Make Decisions and How Those Decisions Go Wrong*. Oxford: Oxford University Press, 2013.

13

Computational Models of Decision-Making

This lecture discusses some exciting theoretical developments in cognitive science that seem to bridge the gap between the brain and behavior. These developments incorporate several ideas that you've explored so far, including normative theories and neural mechanisms of decision-making. Specifically, you'll see how scientists interpret the behavior of decision-making neurons as a kind of Bayesian computation, which has provided insights into why these neurons behave the way they do.

Sequential Probability Ratio Test

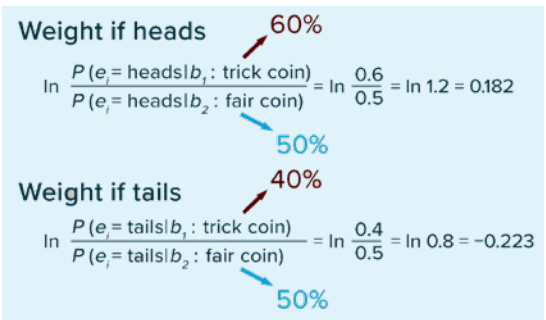
A good place to start is with a discussion of how Bayesian computations can help in decision-making and how they relate to the behavior of decision-making neurons. Suppose you're trying to make a simple perceptual decision, such as whether most of the dots in a group are moving to the left or to the right, and you want to make this decision efficiently. That is, you want to collect a reasonable amount of information, but you don't want to stare at the moving dots forever. Once you're pretty sure of the answer, you want to respond. What's the best way to solve this problem?

In 1945, the Hungarian mathematician Abraham Wald proposed a solution that's now known as the sequential probability ratio test. Here's how it works. Suppose you have two coins. One coin is fair, and the other coin produces a head 60% of the time when it is flipped. You don't know which coin is which, and so you pick one up and start flipping it.

Suppose your first flip produces a head, increasing your belief that it's the trick coin but only by a little bit. After all, the fair coin will also produce a head 50% of the time. So how much should flipping a head increase your belief that it's the trick coin rather than the fair coin?

The sequential probability ratio test provides a mathematical way to think about it. The test takes the ratio of the two separate probabilities and then assigns a weight equal to the natural log of that ratio.

In this case, the trick coin produces a head 60% of the time, while the fair coin produces a head 50% of the time, so the ratio of those probabilities is 0.6 divided by 0.5, which equals 1.2. And the natural logarithm of 1.2 is about 0.18, so the sequential probability ratio test adds 0.18 to its running total of evidence each time you flip a head.



Conversely, suppose you flip a tail. The chance of flipping a tail with the trick coin is only 40%, while the chance of flipping a tail with the fair coin is still 50%. So the ratio this time is 0.4 divided by 0.5, which equals 0.8, and the natural logarithm of that ratio is about –0.22. So the sequential ratio test would subtract 0.22 from the running total of evidence each time you flip a tail.

Suppose that you keep applying that same procedure with each coin flip—adding 0.18 to your running total of evidence each time you flip a head and subtracting 0.22 from the running total each time you flip a tail. When do you stop?

Your stopping criterion will also be the natural logarithm of the ratio of two probabilities. But in this case, it will be the ratio of the minimum accuracy you're willing to accept divided by the maximum percentage of errors you're willing to accept. For example, suppose you want to guarantee that whenever you guess that it's a trick coin, you will be right at least 95% of the time. So the minimum accuracy is 95%, and the maximum error rate is 5%, meaning that the ratio is 0.95 divided by 0.05, which equals 19. The natural log of 19 is around 2.94, and so that should be your stopping criterion. In other words, if your running total of evidence reaches 2.94, then you can stop accumulating evidence and decide that it's probably the trick coin. And you'll be right at least 95% of the time.

Trick coin

$$\frac{\text{minimum accuracy (95\%)}}{\text{maximum errors (5\%)}} = \frac{0.95}{0.05} = 19 \qquad \ln 19 \approx 2.94$$

Fair coin

$$\frac{\text{maximum errors (5\%)}}{\text{minimum accuracy (95\%)}} = \frac{0.05}{0.95} = 0.053 \qquad \ln 0.053 \approx -2.94$$

What does the evidence need to reach for you to decide that you're dealing with the fair coin? In that case, you use the converse ratio, putting the minimum accuracy in the denominator and the maximum error rate in the numerator. You take the log of the ratio of 0.05 divided by 0.95, which equals around 0.053. And the natural log of that is –2.94. So if your running total of evidence reaches –2.94, then you can stop accumulating evidence and decide that it's probably the fair coin. Again, you'll be right at least 95% of the time.

Among all possible sequential tests that achieve the same level of accuracy, this procedure will require the smallest average number of coin flips. Mathematically, it's the most efficient procedure out there.

Evidence Accumulation in Neurons

Many neuroscientists believe that the decision-making neurons that were discussed in lecture 12 perform exactly the kind of probabilistic Bayesian computation demonstrated in the coin example. The idea is that the ramping-up activity observed in the neurons might represent the kind of evidence accumulation seen in the sequential probability ratio test. Activity ramps up as more evidence accumulates in favor of a hypothesis. For example, recall that as the monkey continued to look at a set of dots where most of the dots were moving right, the evidence in favor of the hypothesis of rightward motion grew stronger—and so did the activity in the neurons that were associated with rightward decisions.

Furthermore, when the evidence for rightward motion was stronger, the activity ramped up faster. Conversely, when the evidence was weaker, the activity ramped up slower. That pattern of neuronal activity is also exactly what you'd expect from running the sequential probability test. Stronger evidence also leads to faster accumulation there.

In the coin-flipping example, if the trick coin produced a head 90% of the time rather than just 60% of the time, then each time you flipped a head with that coin, you would increase your running total of evidence by the log of 0.9 divided by 0.5 rather than increasing it by the log of 0.6 divided by 0.5. And since the log of 1.8 is larger than the log of 1.2, you would add more to your running total, and the evidence would accumulate faster.

Weight if heads (90% trick coin)

$$\ln \frac{P(e_i = \text{heads} | b_1 : \text{trick coin})}{P(e_i = \text{heads} | b_2 : \text{fair coin})} = \ln \frac{0.9}{0.5} = \ln 1.8 = 0.588$$

Weight if heads (60% trick coin)

$$\ln \frac{P(e_i = \text{heads} | b_1 : \text{trick coin})}{P(e_i = \text{heads} | b_2 : \text{fair coin})} = \ln \frac{0.6}{0.5} = \ln 1.2 = 0.182$$

But the correspondence goes even further. You derived specific thresholds for when to stop collecting evidence in the sequential probability ratio test. And if the cumulative evidence passed that threshold, then you should stop and commit to the decision, whether it reached the threshold quickly or slowly.

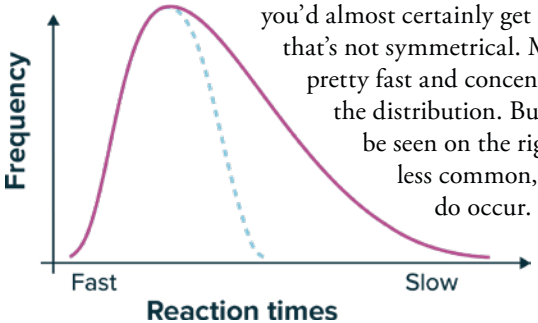
The neurons also behaved that way. Remember, the monkeys always seemed to make their decision about the moving dots only when the firing rate of the decision neurons reached about 65 firings per second. The bottom line is that scientists see this remarkable correspondence between the behavior of neurons involved in decision-making and a well-established Bayesian algorithm that has been proven to be the most efficient algorithm out there.

The Drift Diffusion Model

If you ask people to make simple perceptual decisions, a number of robust empirical phenomena tend to show up. And the same kind of Bayesian computations that explain neural activity also provide very natural accounts of those behavioral phenomena.

One important finding is that if you ask people to respond more quickly, they tend to make more errors, and if you ask them to emphasize accuracy, then their response time tends to slow down. This relationship is known as the speed-accuracy trade-off.

A second very robust finding is that reaction time distributions are positively skewed. For example, imagine that you conduct an experiment where you ask someone to make a bunch of simple perceptual decisions. Some of those decisions are easy, and others are hard, and so you would observe a wide range of reaction times.

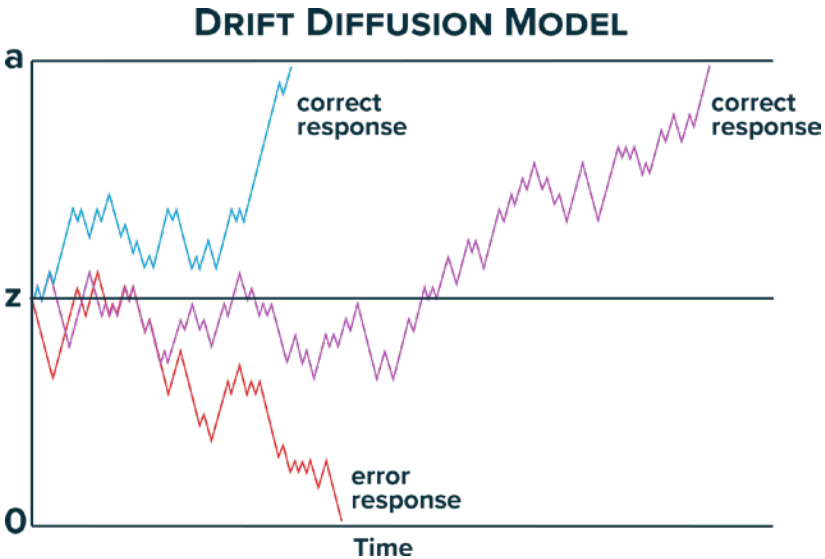


If you plot the frequency of different reaction times, you'd almost certainly get a plot with a distribution that's not symmetrical. Most reaction times are pretty fast and concentrated on the left side of the distribution. But a fairly long tail can also be seen on the right, meaning that, though less common, some long reaction times do occur. Why is that?

Roger Ratcliff at Northwestern University and his colleagues have suggested a computational model that can account for the speed-accuracy trade-off, the asymmetric and positively skewed reaction time distributions, and many more phenomena. It bears a very strong resemblance to the sequential probability ratio test, and it's often referred to as the drift diffusion model.

Time is plotted on the horizontal axis. Plotted on the vertical axis is an arbitrary decision variable based on evidence, where higher values correspond to more evidence for a decision. The top of the graph (*a*) corresponds to the positive threshold for one decision, and the bottom of the graph (0) corresponds to the negative threshold for the other decision.

The three jagged paths in the figure correspond to the evolution of evidence over time during three separate decisions. The two paths that end up crossing the top border would lead to the correct decision, while the path that ends up crossing the bottom border leads to the wrong decision.



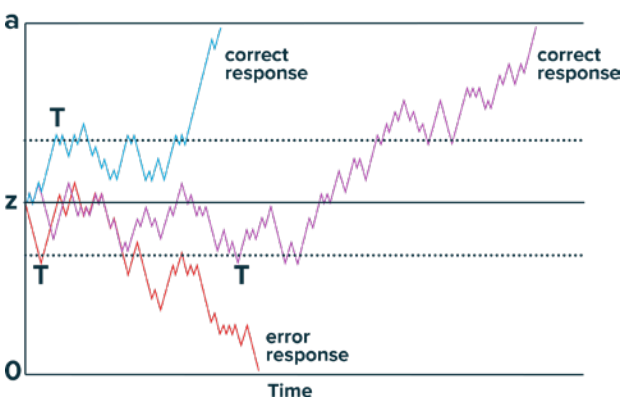
To model decision-making in this drift diffusion model, you first assume that the evidence starts at some specific level (z), and then evidence starts accumulating. If the evidence tends to favor the top response, then the evidence starts moving up, but if it favors the bottom response, then it starts moving down. Finally, when the running total of evidence passes one of the two thresholds, then the model commits and responds with the corresponding decision.

Modeling Decision-Making Phenomena

So how can the drift diffusion model provide insight into the decision-making phenomena discussed in this lecture? First, consider the speed-accuracy trade-off. Remember that the drift diffusion model commits to a decision when the cumulative evidence passes the top or bottom border. So if you want to speed decisions up, just move those borders in closer to the starting point.

That lower threshold is represented by two horizontal lines that are immediately above and below the line corresponding to the starting point. If those lines were the positive and negative thresholds, the evidence wouldn't take as long to cross a border because it doesn't have as far to travel. And that means decisions will be faster.

But it also means that the level of evidence required for a decision is lower and that you'll commit to a decision based on less evidence. And that means that decisions will tend to be less accurate. Graphically, the jagged path of evidence is more likely to cross the wrong border by mistake if that border is closer to the starting point.

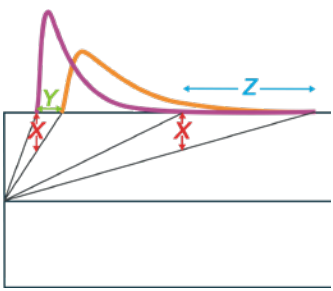


For example, the points in the figure that are labeled with a *T* indicate points when the drift diffusion model would commit to a decision assuming the thresholds had been moved in closer to the starting point. Those decisions get made quickly, but they're more likely to be a mistake. The jagged path in the middle of the figure ultimately passes the threshold at the top corresponding to a correct answer. But with a lower threshold, it crosses the horizontal line just below the starting line. So moving the thresholds closer to the starting point would have led to an error response for that decision. Trying to respond quickly would lead to a mistake.

This process is how the drift diffusion model explains the trade-off between speed and accuracy: You can speed up by requiring less evidence, but doing so makes you less accurate.

The second major phenomenon that was discussed was the fact that reaction time distributions are asymmetric and positively skewed. Most trials lead to fairly fast responses, but a long tail on the right indicates that slow responses also occasionally occur.

The basic idea is that a constant change in the strength of evidence will affect fast decisions differently than slow ones. At the top of the following figure are two distribution curves. They correspond to two sets of response times for two decisions: an easier decision on the left and a harder decision on the right. The straight diagonal lines connect the starting point to the fastest and slowest response times in each of the two distributions—the points at their far ends. These diagonal lines therefore correspond to straight paths of evidence accumulation.



The vertical axis corresponds to evidence, so the two *x*'s correspond to the same constant change in the amount of evidence in different trials. The horizontal axis corresponds to response time. So here, you're comparing the difference in time between the fastest responses for each decision against the difference in time between the slowest responses. Notice that *z* is much larger than *y*.

This graph illustrates that the fastest times for easy and hard trials are pretty similar, and so they tend to bunch up and produce a peak on the left. But the slowest times can be very different, and so they tend to produce a long tail on the right. And according to the drift diffusion model, response time distributions are asymmetric and positively skewed for that reason.

The mathematically motivated Bayesian algorithm can help cognitive scientists understand everything from monkey neurons to human behavior, leading many scientists to the possibility that maybe, at its core, the mind is really a Bayesian computation engine. Rather than dealing in absolutes, maybe the mind's natural currency is really probability.

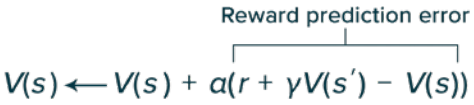
Although not universally accepted, this viewpoint is gaining popularity among cognitive scientists, especially in the area of perception—the idea being that people's perceptions of the world are actually probabilistic hypotheses about what's really out there. But many cognitive scientists are actively exploring how these kinds of Bayesian computations might explain many other aspects of mental function.

The Temporal Difference Algorithm

Another idea in the cognitive science of decision-making has to do with reinforcement learning and the so-called temporal difference algorithm. Recall that reward prediction error signals, which were discussed in lecture 12, can be used to continuously update estimates of the utility or value of different choice options. Why might that happen, and how could it help people learn to make better decisions?

According to reinforcement learning theory, people are constantly estimating the value of different situations in the world. And they can then use those value estimates to choose actions that lead to higher value states. But how do people estimate the long-term value of a situation as opposed to its immediate reward?

That step is where reward prediction errors and the temporal difference learning algorithm come in. The basic idea is summarized in this formula:



V of *s* is your current estimate of the long-term value of state *s*. Likewise, *V* of *s* prime (s′) is your current estimate of the long-term value of state *s*′, where *s*′ is the next state you are in after state *s*. The *r* represents the immediate reward, if any, that you receive from moving into state *s*′. Alpha (α) is a small fraction that represents the learning rate, or how fast you should change your estimates of long-term value after each state transition. And gamma (γ) is a fraction called a discount factor, which allows you to discount the value of future states relative to current states. For this purpose, you can assume gamma is 1 and ignore it.

The assumption is that the long-term value of your current state should equal the long-term value of the next state you enter plus any immediate reward that you receive. But initially, your value estimates are probably not very good, and so they need to get updated. And that procedure is what the temporal difference rule in this formula does. Specifically, after you've reached the next state *s*′, it updates your estimate of the value of your previous state *s*. That updated estimate is what it spits out on the left side of the arrow. And it does that using the reward prediction error.

On the right side of the formula, the first two terms in the parentheses are the sum of the immediate reward, *r*, plus your estimate of the long-term value of the next state *s*′. And that sum should be equal to your estimate of the long-term value of the previous state *s*. If it isn't, then you've got a reward prediction error. You either underestimated or overestimated the long-term value of the previous state. So this formula will make a slight change to your value estimate of the previous state every time you change states.

For example, suppose you're playing chess, and you think you're losing. But then, after you make a move, you realize that the situation is better than you thought, and you're actually likely to win. In that case, you would increase your estimate of the long-term value of the state you were in before the move. Conversely, if you thought your chances of winning got worse after the move, then you would decrease your estimate of the value of the earlier state.

That algorithm is extremely powerful. In fact, it forms the basis for some of the most impressive systems in artificial intelligence today, some of which you'll learn about later.

Reading

Dayan, P., and L. F. Abbott. "Classical Conditioning and Reinforcement Learning." In *Theoretical Neuroscience*, 331–339. Cambridge, MA: MIT Press, 2001.

Sutton, R. S., and A. G. Barto. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 2018.

Wald, A. "Sequential Tests of Statistical Hypotheses." In *Breakthroughs in Statistics: Foundations and Basic Theory*, 256–298. New York: Springer, 1992.

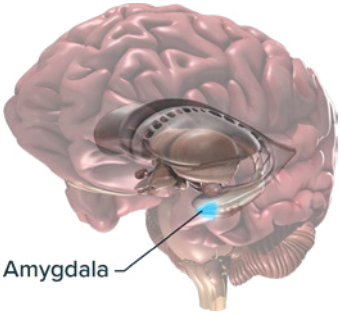
Wang, Z. J., and J. R. Busemeyer. *Cognitive Choice Modeling*. Cambridge, MA: MIT Press, 2021.

14

The Emotional Brain

For a long time, much of the research in cognitive science ignored emotions, even though scientists in the field realized that emotions play an enormous role in human life. One reason for such treatment might be that emotions are felt subjectively, and it's tough to study subjective experience using objective scientific methods. How do you measure a feeling? And if you can't even measure it, how can you study it scientifically? But over the past few decades, cognitive science has made some real progress in understanding emotion, including the neural mechanisms involved. This lecture looks at several theories and studies that have shed light on this elusive topic.

A Case of Fearlessness



A neurological patient named S. M.—referred to by her initials to protect her identity—suffers from an extremely rare genetic disorder called Urbach-Wiethe disease that leads to the progressive deterioration of a brain structure called the amygdala in both hemispheres. This disease also produces an extremely rare symptom: In most situations, S. M. seems to be completely and utterly fearless.

For example, consider Waverly Hills Sanatorium in Louisville, Kentucky. This abandoned building regularly serves as a haunted house around Halloween and is considered to be among the scariest of such attractions in the country. S. M. walked through it with amusement and excitement but no trace of fear. Her companions, however, were terrified.

Likewise, she had no qualms about handling snakes and spiders in an exotic pet store. She also apparently can't recognize fear in other people or even draw a scared face, though she can recognize and draw other emotions, such as happiness.

S. M.'s lack of fear has put her in many dangerous situations. As a result, she has experienced a lot of traumatic events in her life. In each situation, she never exhibited the normal fight-or-flight response, and the experiences didn't lead her to act more cautiously in the future.

Other aspects of S. M.'s mental life are quite normal, including her memory, language skills, vision, hearing, and IQ. Even her experience of other emotions, positive and negative, seems to be normal.

Amazingly, S. M. is immune only to fear of threats from her environment, or what is called exteroceptive fear. In contrast, she has a normal or perhaps exaggerated fear response to interoceptive threats coming from her own body. For example, after a very painful dental procedure, S. M. developed a pathological fear of the dentist. She never saw a dentist again despite having significant dental problems. And unfortunately, she lost all of her teeth.

Studies of patient S. M. suggest a number of important hypotheses about emotion and its neural substrates. First, they suggest that different emotions are quite distinct from each other, not only in how they feel but also in the neural circuits involved. The amygdala, which is the part of the brain that was damaged in S. M., seems to play an important role in processing fear, but the fact that S. M.'s other emotions are relatively normal suggests it's less important for other emotions.

Second, fear itself can be further subdivided into exteroceptive and interoceptive fear. Some scientists have suggested that the amygdala is particularly involved in exteroceptive fear and is much less involved in interoceptive fear.

Third, S. M.'s case also demonstrates the functional utility of emotions, even negative emotions like fear. They motivate people to adapt their behavior in ways that increase their chances of surviving and thriving.

The James-Lange Theory

Many theories attempt to explain what emotions are, why humans experience them, and what function they serve. One of the first theories of emotion is often referred to as the James-Lange theory, after the famous 19th-century scientists William James and Carl Lange, who independently proposed similar ideas in the 1880s.

The basic idea behind the theory is that emotions correspond to people's subjective experiences of underlying physiological events. Many people intuitively believe that their evaluation of some situation makes them happy or sad or angry and that the emotion then leads to physiological changes, such as increased heart rate or rapid breathing. The James-Lange theory assumes the opposite: that the physiological changes themselves cause emotion.

According to this view, the feeling of fear is based on a racing heartbeat and a dry mouth. And if you take away the underlying physiological changes, then you take away the experience of the emotion itself.

One prediction of this theory is that you could potentially influence your mood by making physiological changes. For example, smiling might make you feel happier, and frowning might make you feel sadder. This outcome is sometimes referred to as the facial-feedback effect, and evidence to support it indeed exists.

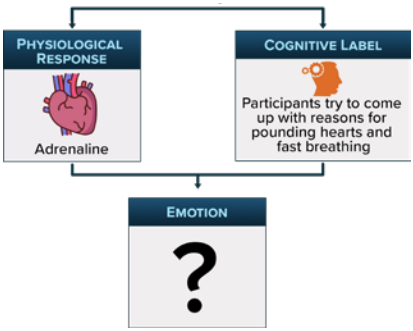
In 2022, a large multinational study of the effect tested nearly 4,000 people in 19 different countries. Scientists found that smiling, as well as mimicking the smile of someone else, could amplify and even initiate feelings of happiness.

But Harvard physiologist Walter Cannon and his graduate student Philip Bard discovered a major problem with the theory. They showed that emotions don't always depend on physiological changes, as the model assumes. Consider a quadriplegic whose spinal cord has been completely severed and whose brain therefore does not receive any signals from the body. If emotions depend on physiological changes in the body, then one would expect that such a person might not experience emotions at all or, at the very least, that their emotional experience would be dramatically different than that of other people. But such people do experience the full range of emotions, just like everyone else. Cannon and Bard therefore argued that emotions and physiological changes are independent of each other and that emotions can occur without physiological changes.

The Two-Factor Theory

In the 1960s, Stanley Schachter and Jerome Singer proposed a two-factor theory of emotion that incorporates both physiological and cognitive mechanisms. For example, say you see a snake, and, instinctively, your heart starts pounding, and you begin breathing faster. That mechanism is factor one: the physiological response.

But then you immediately try to explain why your heart is pounding and why you're breathing fast. And seeing the potentially dangerous snake allows you to attribute the physiological response to fear of the snake. That mechanism is factor two: the cognitive label.



So even though emotions often start with a physiological response, they actually reflect a cognitive appraisal. According to the two-factor theory, the same physiological response could lead to completely different emotions if the cognitive appraisal is different.

Schachter and Singer conducted a famous but controversial experiment that provided the original evidence for their two-factor theory. Participants were given an injection of what they were told was a drug called Suproxin. But they were actually injected with either adrenaline or a placebo that should have no effect. Some of the participants were told that the side effects of Suproxin were similar to the effects of adrenaline: a pounding heart, shaking hands, and a flushed face. That was the so-called informed group.

Other participants were misinformed about the effects of the drug. They were told that they might experience a headache and some itching and that their feet might go numb. That was the misinformed group.

Still other participants who were injected with adrenaline weren't told anything about what to expect. That was the ignorant group. And finally, the people in the control group were injected with the placebo and also weren't told anything about what side effects to expect.

INFORMED GROUP
Told that side effects were similar to the effects of adrenaline

MISINFORMED GROUP
Told that side effects included headaches, itching, numbness in feet

IGNORANT GROUP
Not told any side effects

CONTROL GROUP
Received a placebo and not told any side effects

All the participants were told to wait in a room with another participant while the drug kicked in. But the other person in the room was actually a confederate of the experimenters who acted either happy or angry. The experimenters watched the interactions and rated the participants' behavior against some objective criteria that would typically be associated with happiness or with anger. They also gave the participants a questionnaire and asked them to rate how happy they felt and how irritated or angry they felt.

They found that if the participants expected the physiological response based on the injection, then they didn't look any further for an explanation. But if they weren't expecting the injection to produce the physiological response they experienced, then they did look further. And if they saw a confederate who seemed happy, they mistakenly attributed their arousal to being happy. Conversely, if they saw a confederate who seemed irritated, they mistakenly attributed their arousal to being angry. And as a result, the people who didn't expect the drug to make them feel aroused reported feeling happier and angrier than the people who could attribute their arousal to the injection.

Unfortunately, the original experiment has been difficult to replicate, and so it's unclear whether the reported effects are reliable or not. Nevertheless, the two-factor theory has played an extremely important role in the field.

Facial Expressions and Autonomic Signals

One interesting assumption of the two-factor theory is also a matter of significant debate: Can the same underlying physiological changes produce two different emotions? Or does each emotion have its own physiological signature? For example, do different neural circuits in your brain get activated depending on whether you're afraid, surprised, or happy?

A number of emotion researchers have argued for such a view and have collected data that seems to support it. In one very famous set of studies, Paul Ekman and his colleagues asked people from a wide variety of cultures to pick out a face that conveyed a specific emotion.

The original studies tested pictures that were meant to convey six different emotions: happiness, sadness, fear, anger, surprise, and disgust. And what these studies found was that people in different countries all tended to pick out the same faces when asked about each of the emotions.

The researchers concluded that the association between a specific emotion and a specific facial expression was universal. Facial expressions are a kind of signature or fingerprint that can be used to determine what specific emotion is being experienced.

Ekman went on to test whether activity in the autonomic nervous system—which is responsible for things like heart rate and breathing—can also distinguish between different emotions. Are anger, sadness, and disgust associated with different autonomic signatures, as well as facial expressions? This study was done in collaboration with Robert Levenson and Wallace Friesen at the University of California, San Francisco.

The results suggested that specific emotions were indeed associated with specific patterns of autonomic activity. For example, anger, fear, and sadness were all associated with a higher heart rate than happiness, disgust, and surprise. And anger was associated with a higher skin temperature than any of the other emotions.

So each emotion seems to have its own telltale signature in the human body. Although some emotion scientists find these results convincing, others do not. Skeptics point out that the studies of facial expressions were subjective and involved human judgment. They also typically involved posed faces that never moved. Furthermore, when scientists try to do more objective studies, the results don't typically turn out the same way.

To be sure, most people do associate wide eyes with fear and frowning with sadness. But the skeptics would claim that those are just learned stereotypes and that they don't constitute a reliable signature of emotion. In reality, people exhibit a huge variety of different facial expressions when they're afraid and not just one.

But what about the study that measured autonomic signals? Critics point out a couple of potential concerns. For one, with the exception of anger, the emotions did not elicit a unique autonomic signature. For example,

fear and sadness were similar. And so were happiness, surprise, and disgust. Furthermore, subsequent experiments that also looked for autonomic signatures of specific emotions often failed to replicate the original results and sometimes even associated the same emotion with a fundamentally different autonomic pattern than Ekman and colleagues had found.

Constructionist Theories

Lisa Barrett at Northeastern University proposed an alternative constructionist theory of emotion. Rather than assuming that emotions are genetically hardwired from birth and that unique physiological and autonomic signatures of each core emotion exist, Barrett proposed that emotions are constructed on the fly by categorizing real-time experience. She put it this way:

> Emotions are not triggered; you create them. They emerge as a combination of the physical properties of your body, a flexible brain that wires itself to whatever environment it develops in, and your culture and upbringing, which provide that environment.

Color is sometimes used as an analogy. When you look out at the world, you categorize the colors that you see into discrete groups: blue, red, green, and so on. But of course, no discontinuities exist in the wavelengths of light that you see. Rather, you're imposing a specific set of categories onto the wavelengths that you see.

Constructionist theories argue that the same thing is going on with emotions. When people sort and label their feelings, they are imposing a discrete set of emotional categories. But the underlying physiological input is in fact multidimensional and continuous. And the way that people break up that input into distinct categories might change depending on the context.

So can this kind of constructionist theory be distinguished from the more traditional theory that different emotions depend on specific neural circuits? One approach is to analyze neuroimaging studies that investigated emotion. Are the same brain regions consistently activated by a specific emotion, and are those activations specific to that emotion?

Kristen Lindquist at Harvard Medical School, Lisa Barrett, and other scientists analyzed results from 91 different neuroimaging studies of emotion. Their review failed to find clear evidence that specific emotions were localized to specific brain regions. Although some brain regions did exhibit consistent activation during specific emotions, that activation was not specific to that emotion.

The researchers concluded that the neuroimaging evidence was more consistent with a constructionist theory of emotion than with a traditional locationist theory. Different emotions did not correspond to certain local circuits in the brain. Rather, the same neural pathways were often active while experiencing vastly different emotions.

The take-home message is that emotions don't map one to one onto specific neural circuits or specific autonomic signatures. Instead, you should look at emotions as mental constructions. They can certainly be influenced by autonomic responses, but they also depend on contextual factors and how you interpret any given situation.

Reading

Amaral, D. G., and R. Adolphs, eds. *Living without an Amygdala*. New Yok: Guilford Press, 2016.

Barrett, L. F. *How Emotions Are Made: The Secret Life of the Brain*. London: Macmillan, 2017.

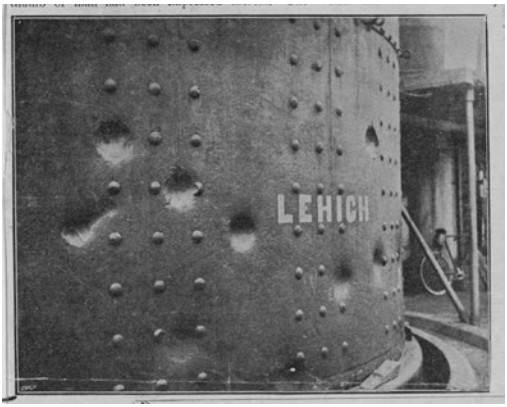
LeDoux, J. *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. New York: Simon & Schuster, 1998.

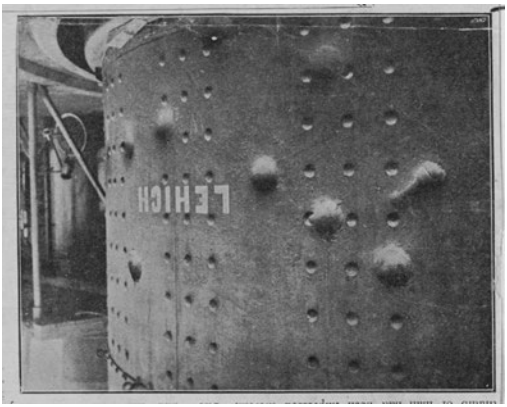
15

The Science of Perception and Illusion

This lecture dives into a topic that has received a lot of attention in cognitive science: visual perception. How does your mind take in the vast quantity of information that is constantly impinging on your eyeballs and then produce an accurate representation of what's out there in the world? Cognitive science has made substantial progress in answering that question.

Perceiving Light and Shadow





This close-up picture is of a Civil War–era warship, the USS *Lehigh*, with a bunch of dents going in and bumps coming out. Which are bigger, the dents or the bumps? Virtually everyone would agree that the dents going in are bigger than the bumps coming out.

But when you rotate the picture 180°, do you see the dents change into bumps? What's going on?

Notice the shadows. Your perceptual system knows that light usually comes from above, and it uses that fact to help it perceive depth. Assuming that light comes from above, the big circles in the original picture have to be going in because the shadow is on top, and reflected light is on the bottom. The inverse is true when the picture is inversed.

Foundations of Perception

Perception can be defined in different ways, but here's one helpful definition: Perception refers to the mechanisms you use to take in information from the environment via your sense organs and transform that information into internal representations of what you see, hear, taste, touch, and so on. And you can use that representation for further cognitive processing. But how do you go from what comes into your sense organs to an accurate internal representation?

Take a look at this picture. How do you go from a pattern of light that happens to reach your retina to an internal representation of a group of roughly spherical objects that are red and dimpled and have some green things sticking out of the top? That problem is what perception is trying to solve.

You also want to recognize what you perceive by matching it against representations that you have stored in your memory. But before you can do that, you first have to be able to build a representation of the visual input, which is what perception is supposed to do.

When cognitive scientists study perception, they distinguish between the distal stimulus, the proximal stimulus, and the internal representation. The distal stimulus is the object out in the world that you are trying to perceive—the thing at a distance from you, hence the term *distal stimulus*. So if you look around your home and see a table, then the real table out in the real world is the distal stimulus.

And if your perceptual system works, then you should be able to say things like, "Oh, that's a rectangular thing with vertical things coming down, and it's brown and looks smooth."

The term *proximal* means "close." So the proximal stimulus is close to you. In particular, it's the pattern that the distal stimulus makes on your sensory organs. For example, when you look at the table in your home, it creates an image on your two retinas, and that image is the proximal stimulus.

Similarly, when you listen to someone speak, the sounds cause your ear drums to vibrate in specific ways, which is the proximal stimulus. In the case of touch, the proximal stimulus would be the pattern of indentations on your skin.

Ultimately, you want to derive an internal representation of the distal stimulus that describes what you're perceiving. For example, if you're looking at a table, the internal representation should indicate that it's rectangular, that it's brown, that it's smooth, that it's about 3 or 4 feet high, and so on. If you're perceiving a bird's song, your internal representation should reflect the pitch or frequency of the sounds and the order that they occur in. If you're touching a surface, it should tell you whether the surface is smooth or rough, hot or cold, soft or hard, and so on.

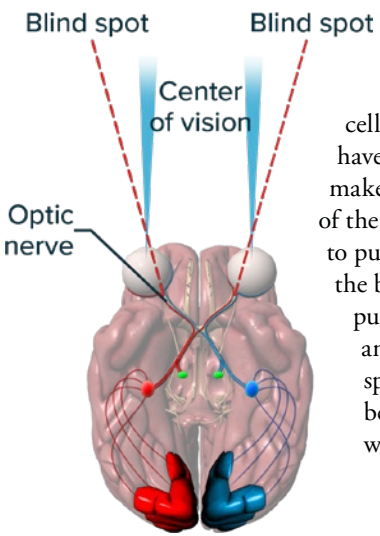
Keep in mind that people can, and regularly do, perceive things that they don't recognize. For example, if you hear an unfamiliar sound like "selimut," you can perceive that sound and create an internal representation of it that is accurate enough to allow you to repeat it even if you don't recognize it as the Indonesian word for "blanket."

So perception refers to the mechanisms used to transform the proximal stimulus into an internal representation. Note that perception is not about converting the distal stimulus into the proximal stimulus. That realm is physics, like optics and acoustics—not cognitive science.

Perception starts at the proximal stimulus. How do you go from the pattern of light on your retinas or the pattern of sound waves on your ear drums to an internal representation of what you're seeing or hearing? That domain is perception.

Major Characteristics of Perception

A common belief is that perception always directly delivers an accurate representation of the external environment. You might think that when you look around at the world, you always see it the way it actually is, but that's not the case. The truth is that your perceptual system only encodes a very limited amount of information about the environment at any given time, and then it makes a guess about what's actually out there based on that information.

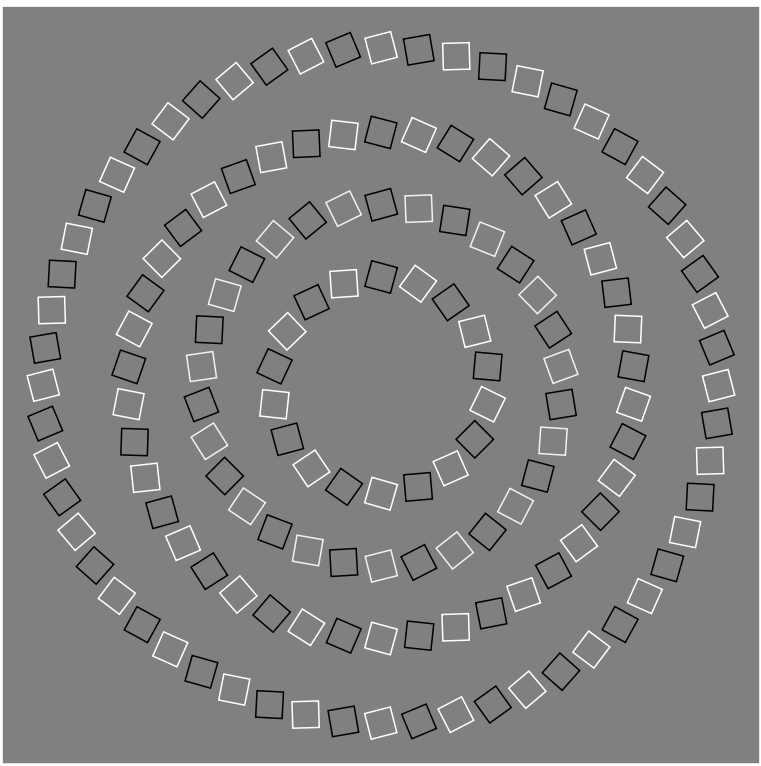


Consider, for example, that you have a blind spot in both of your eyes—a spot on your retina that doesn't contain any photoreceptors, the rod and cone cells that respond to light. The reason you have those blind spots is because the fibers that make up the optic nerve originate on the front of the retina, and so the optic nerve actually has to punch through the retina to head back into the brain. And the place where the optic nerve punches through is a hole that doesn't contain any photoreceptors. So you're blind in that spot. You're never aware of your blind spot because your brain always fills in the gap with an appropriate guess.

And it turns out that your brain does this all the time. For example, you have high-resolution vision only in the 1% of the visual field that your eyes are focused on. Your peripheral vision is much worse, but you have the impression that you see everything clearly. In reality, you need to look directly at something to see it clearly, and your brain fills in the periphery with a plausible guess.

Of course, your brain is extremely good at making those guesses, but it does occasionally make mistakes, which is when you experience a perceptual illusion: What you think you see is not actually what's out there. The internal representation and the distal stimulus don't match.

For example, take a look at this image.



When most people look at this picture, they think they see spirals that cross over each other. But the picture is actually just four concentric circles that vary in size without any spiraling, and none of the circles actually cross over each other. But this picture fools your perceptual system and leads you to create an internal representation that doesn't correspond to the actual distal stimuli.

Another very interesting characteristic of the human perceptual system is what cognitive scientists sometimes refer to as paradoxical correspondence. Put simply, people often construct accurate internal representations of the distal stimulus even when the information available in the proximal stimulus is radically incomplete or misleading.

One of the best examples of paradoxical correspondence is depth perception. People typically perceive depth so accurately and effortlessly that it's easy to forget how almost miraculous this ability is.

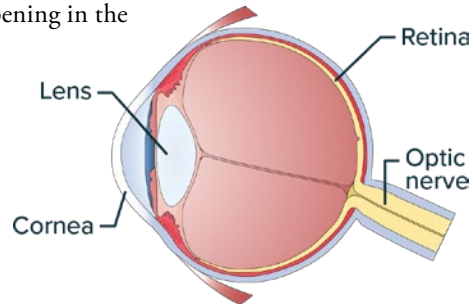
The real world is three-dimensional, but your retinas are two-dimensional and flat. So the proximal stimulus is just a flat image of the world. That image is all your brain gets. Your brain doesn't have direct access to the distal stimulus in the real world. So how can you see depth given that the input is flat?

That contradiction is an example of a paradoxical correspondence. Your internal representation of the world is three-dimensional, and that does correspond to the distal stimulus. But the correspondence is paradoxical because the proximal stimulus is flat. In this case, your perceptual system is also making a guess based on incomplete information, but, fortunately, it's extremely good at making those guesses.

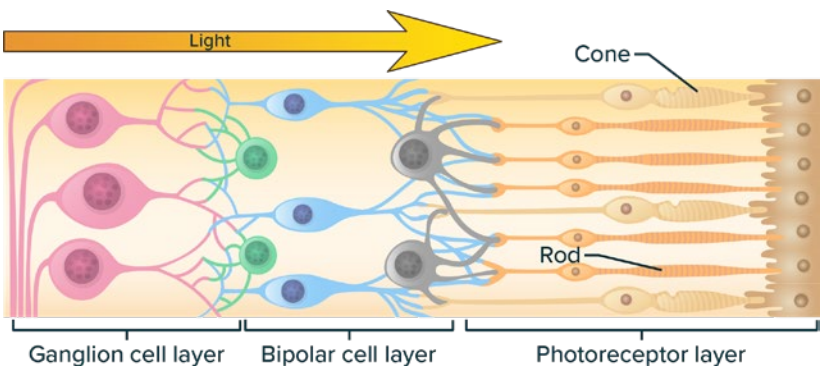
Your perceptual system doesn't always encode the world the way it really is. Instead, it takes a limited amount of sensory information and then makes a guess about what's really out there. And although it's very good at creating an internal representation based on a guess, sometimes it's wrong.

How the Eye Processes Visual Information

The eye is a good place to start if you want to understand what's going on in your brain when you process a visual image. On the back of each eyeball, you've got a sheet of cells called the retina, which is actually part of your nervous system. And a lot of information processing is happening in the retina itself. If you examined a small part of the retina under a microscope, you'd see layers of cells.



On the surface of the retina are the ganglion cells. Behind them is the bipolar cell layer. On the back of the retina are the photoreceptor cells—the cells that actually respond to light. So the light comes in through the eye and has to get through the ganglion cells and the bipolar cell layer before it gets to the photoreceptors in the back.



You have two different kinds of photoreceptors: rods and cones. The rods detect brightness—how light or dark something is—and they can work with very little light. The cones detect color, and they need a fair bit of light to work. The cones are concentrated in the fovea, that part of the retina that corresponds to central vision—where you're looking. The rest of your retina is dominated by rods, which play a major role in your peripheral vision.

This organization actually has some implications in your daily life. For example, you may have noticed that when it's dark, you don't see colors very vividly. The reason is that your cones aren't getting enough light to fire, and your perception is depending mainly on your rods. But your rods are color blind.

Likewise, you may have noticed that your peripheral vision is typically better than your central vision when it's dark. The reason is that the cones in your fovea need a fair bit of light, so they'll be pretty useless in the dark. Alternatively, the rods in your peripheral vision don't need nearly as much light, so they'll respond better.

After the photoreceptors have processed the incoming information, they send their signals forward in the retina through the bipolar cell layer to the ganglion cells, which are the output cells of the retina. The ganglion cells do some further processing of the visual information and pass on their outputs to the thalamus, which is like the brain's switchboard. Most information coming from all the senses gets sent to the thalamus first before moving on to the cerebral cortex for further processing.

How the Brain Processes Visual Information

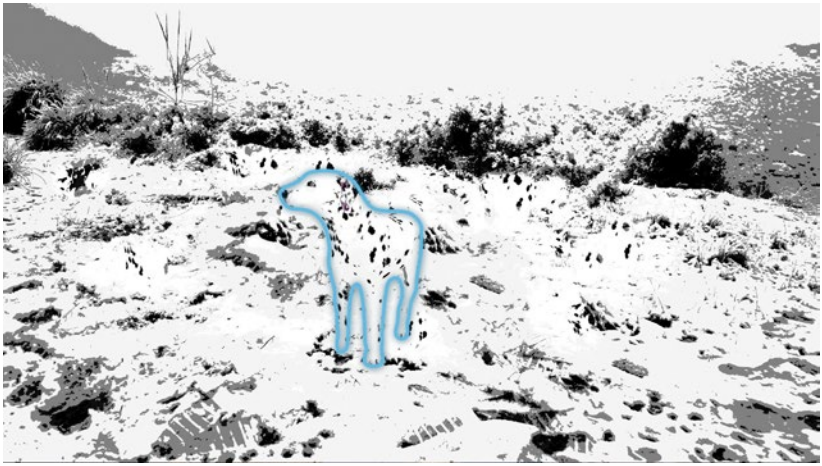
And at each processing stage, the brain extracts more and more complex features from the visual input. The ganglion cells in the retina just recognize dots of light at different locations in the visual field and send that information to the brain. Cells in the visual cortex of the brain process the low-level information they receive and begin to recognize lines and edges.

Those cells feed into yet other cells that begin to recognize more complex features, like moving lines, corners, and gaps. This process continues through multiple processing layers until finally reaching cells that recognize more complete objects.

Notice that the brain can't just directly perceive complete objects at a glance, even if it seems that way to you. Instead it has to construct internal representations of objects feature by feature from the bottom up. Given that process, all the information might seem to flow in one direction from the bottom up. But it turns out that what you perceive can also be strongly influenced by what you expect and know about the world.

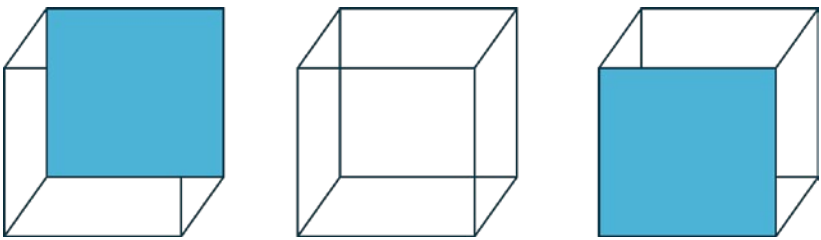
In some cases, changing a person's expectations can radically and permanently change their perception. For example, when you see this picture for the first time, it probably looks like just a bunch of random splotches.

But when you're given an expectation of what it actually is, then you can see that it's a Dalmatian.



Now if you go back to the previous image, you probably can't help but see the Dalmatian. Because you know what to expect, you can't see it as just splotches anymore. Your perception isn't just based on what's coming in from your sensory organs; your knowledge and expectations are also having a profound influence on what you perceive.

Here's another example that also illustrates how your expectations can influence your perception: a so-called Necker cube. The cube can be perceived in two contradictory ways. You can view the top-right square as being in the front (the image on the left), or you can view the bottom-left square as being in the front (the image on the right).



Many people can actually choose which way to perceive it and switch back and forth. Something must be inside their heads that allows them to construct the internal representation that they want. But notice that you can't see it both ways simultaneously. Your internal representation changes based on which square you expect to be in the front. So once again, you can see how your knowledge and expectations have a profound impact on your perception.

Reading

Breitmeyer, B. *Blindspots: The Many Ways We Cannot See*. Oxford: Oxford University Press, 2010.

Goldstein, E. B., and L. Cacciamani. *Sensation and Perception*. Boston: Cengage, 2021.

Tovée, M. J. *An Introduction to the Visual System*. Cambridge: Cambridge University Press, 1996.

16

Computational Models of Vision

In the last couple of decades, cognitive scientists have made exciting progress in their efforts to figure out how cognitive systems like perception work at a mechanistic level. This lecture focuses on one of those areas of progress—mechanistic models of visual processing—and, specifically, models of object recognition.

Simple and Complex Cells

Object recognition is how you identify and categorize objects in your visual environment. It involves more than just perceiving visual features. You also need to figure out what objects those visual features correspond to. Are you looking at a chair, a car, a dog, or a cloud?

Researchers have proposed several computational models of object recognition, and some of the most successful ones are based on deep neural networks. Many modern models incorporate findings from neuroscientific studies that were conducted by David Hubel and Torsten Wiesel in the 1960s.

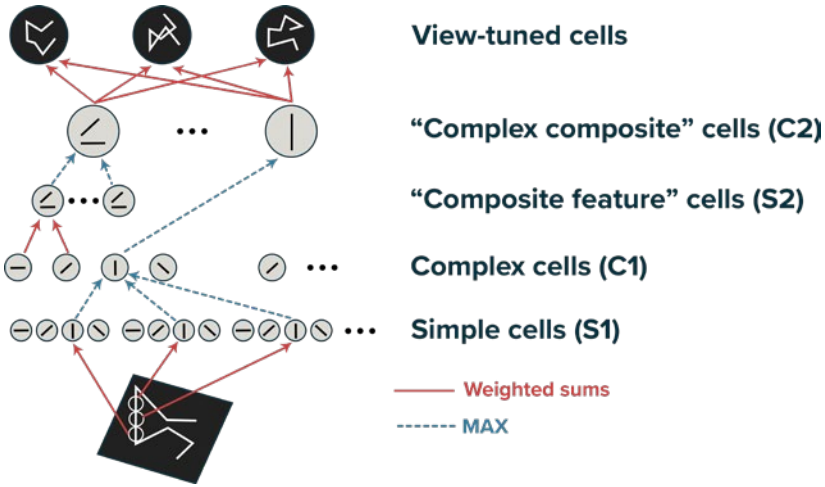
Hubel and Wiesel studied the responses of individual neurons in the visual cortex of cats. They found that many neurons responded selectively to particular features of the visual environment, such as edges or lines of a specific orientation. For example, a particular neuron in the visual cortex always fired in response to vertical bars of light. Another cell fired in response to horizontal light bars, and another one fired in response to light bars at a certain diagonal orientation. Hubel and Wiesel called cells like this simple cells because they respond only to a specific arrangement of light and dark regions within a particular part of the visual field.

The scientists distinguished the simple cells from what they called complex cells, which they hypothesized responded to patterns of activity across sets of simple cells. And this kind of arrangement allows complex cells to respond to more complex features, like moving edges and particular configurations of lines.

The HMAX Model

The hierarchical max (HMAX) model is one of the most influential computational models of the brain's visual system. It took a lot of inspiration from the work of Hubel and Wiesel, particularly the distinction between simple cells and complex cells.

So, what happens when scientists show an image to HMAX? Just like simple cells, the first layer of cells will detect features like edges and bars of light in specific orientations and in specific parts of the visual field.

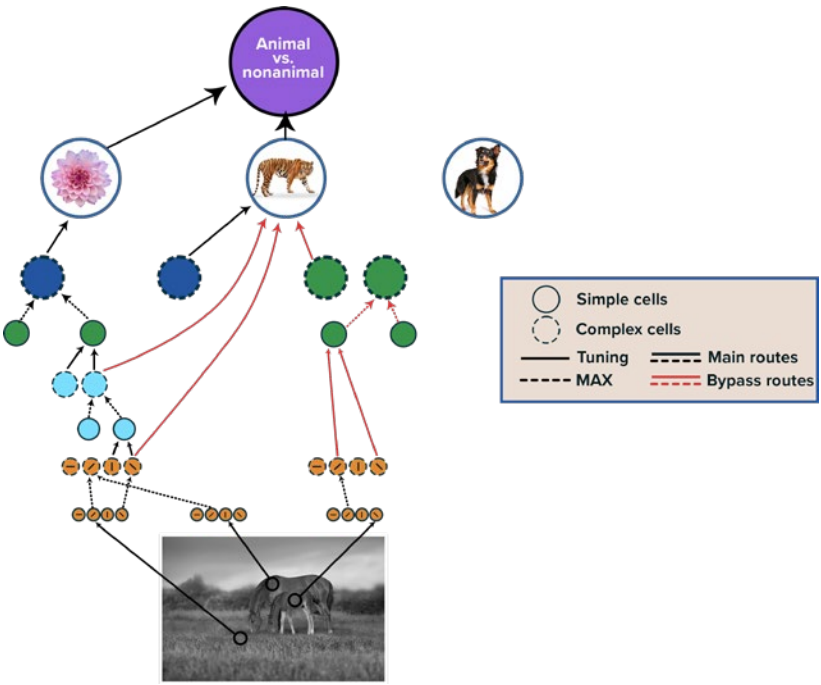


The second layer of cells are like complex cells. They respond to patterns of activity across the simple cells. In particular, they pool together the outputs of simple cells that respond to similar features but in slightly different parts of the visual field. For example, suppose you've got five simple cells that all respond to vertical bars, but each one responds to vertical bars in nearby but different parts of the visual field. Then, you might have a complex cell in the next layer that responds when any of those five cells are firing. That's often referred to as max pooling because you pool across a set of simple cells and respond based on the maximum activity across all of them.

And this kind of max pooling operation allows the HMAX model to recognize a feature even if it's slightly shifted in space. In other words, it allows HMAX to achieve spatial invariance: It doesn't matter if you see a letter *T* in slightly different locations on a page. You can recognize it as a *T* regardless of its spatial location.

Pooling together multiple simple cells also expands the visual field for complex cells. Neuroscientists typically refer to this as the size of a neuron's receptive field. Simple cells have relatively small receptive fields, while complex cells have slightly larger receptive fields.

Variants of the HMAX model have included multiple layers of simple and complex cells in a hierarchy in an attempt to match the hierarchy of visual regions in the brain. For example, Thomas Serre and his colleagues at MIT built a computational model with four sets of alternating cell layers: simple, then complex, then simple, and then complex again. Each layer mapped onto a specific part of the visual system.

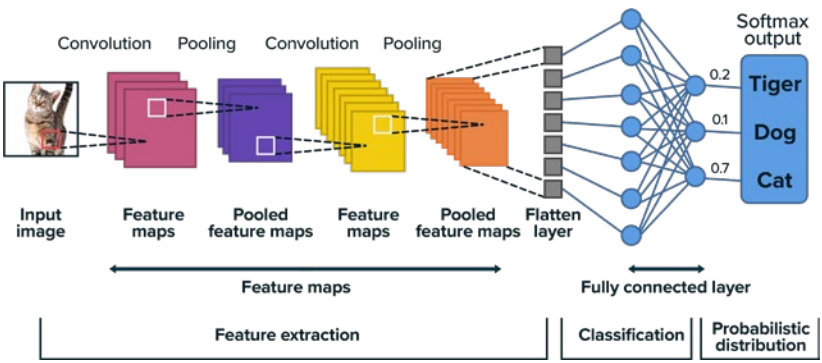


As you move through the hierarchy, the max pooling operation of the complex cells makes it more and more invariant to spatial location: The model can recognize features across an increasingly wide visual field. And the simple cells that sit higher in the hierarchy are receiving the input of complex cells lower in the hierarchy. That means these simple cells can recognize more and more complex combinations of features and increasingly complex visual stimuli.

Then, the next layer of complex cells will pool their input again, and the process continues. As information is processed through the layers, the model becomes able to recognize increasingly complex objects and can perform real visual tasks, like distinguishing animals from nonanimals.

Convolutional Neural Networks

Neurally inspired models like HMAX and its successors inspired researchers in artificial intelligence to build something that has since revolutionized computer vision: a convolutional neural network, or CNN. Like the HMAX model, CNNs consist of a series of processing layers that extract increasingly complex and invariant features from the input data.

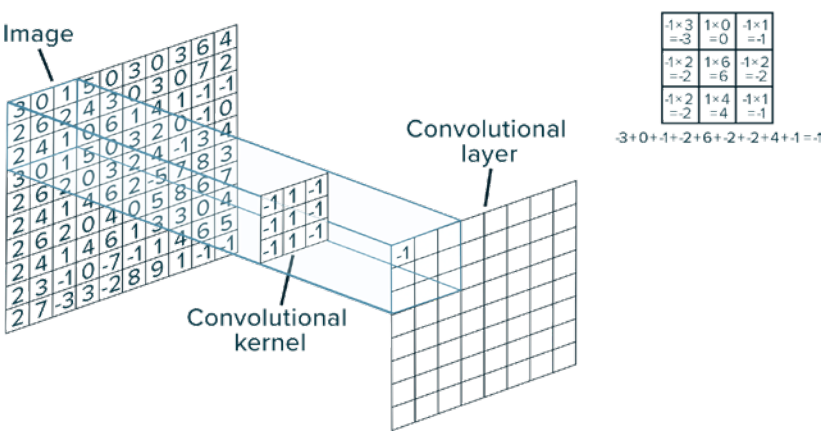


CNNs also use alternating sets of layers that are similar to simple cells and complex cells. One set of layers recognizes specific visual features in specific parts of the visual field, and then the next set of layers pools across different parts of the visual field so that the model becomes spatially invariant. Then, the next set of layers recognizes even more complex visual features, and the layers after that pool across spatial locations again. By the time you get to the highest layers, the CNN can recognize very complex features, like a car or a table, and it can do so regardless of where the object appears in the visual field.

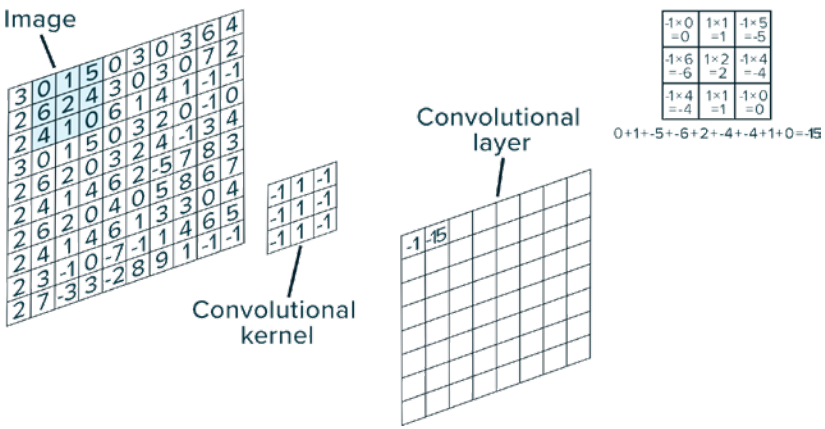
Like other deep neural networks, CNNs use layers of artificial neurons. But CNNs uniquely use convolutional layers, which are like filters or kernels that slide over each location of the input image. At every location, the layer will perform a mathematical operation known as convolution, which starts with two sets of values: the values of the filter and the values of the data from the input image. The convolution operation then multiplies the values of the filter with the corresponding values of the input data and sums up the results.

For example, imagine a three-by-three convolutional kernel as it slides over an image that is much bigger. In this scenario, the convolutional kernel has three columns, and the leftmost and rightmost columns consist of –1s, and the middle column consists of +1s. Such a kernel will identify parts of the image where there is a vertical edge of light surrounded by darkness on the left and right.

If you position the kernel at the top-left corner of the image, it sits atop a three-by-three piece of the image. Then, each of the nine values in that three-by-three kernel are multiplied element-by-element with the nine pixel values directly beneath the kernel. The resulting products are added together to get a single value, and that value is then passed on to the top-left neuron in the next layer of cells. That next layer is called a convolutional layer because its activity reflects the output of a convolution operation.



The kernel is then moved one pixel to the right, and the process is repeated on the next three-by-three region of the image, and then the next region, until the kernel has covered the entire row. Next, the kernel is moved down to the second row, and the process is repeated again. And then, it does the same on the third row, fourth row, and so on, until the entire image has been covered. This results in an activation pattern on the next layer of cells that represents the convolution of the kernel and the original image.



To spark the most activity in that next convolutional layer, you want the sum of products to be as big as possible. Since the kernel multiplies the left and right columns by –1 and the middle column by +1, the largest sum of products will be in parts of the image where the middle pixels are very positive. Typically, larger values correspond to brighter parts of the image. And so, wherever there's a vertical edge of light with darkness to the left and right, the convolution is going to produce a large value.

And voila! You have a vertical edge detector. In fact, the next layer of cells will actually be a map indicating where there are bright vertical bars in the image. If there's a vertical bar in the top left and another in the bottom right, you'll have lots of activity in those parts of the map and lower activity elsewhere. These kinds of maps are often called feature maps because they indicate where a specific feature appears in the image.

And CNNs typically involve lots of different convolutional kernels leading to lots of different feature maps. One map might indicate the location of vertical bars of light, while another indicates the location of horizontal dark bars. And yet another indicates the location of diagonal bars.

In many ways, these convolutional layers are a lot like layers of simple cells in HMAX. But there's a key difference. The kernels used in convolutional layers are actually learned during the training process. Specifically, the neural network is presented with tons of images along with the correct classification for each image. And with each image, the network adjusts its synaptic weights using the backpropagation algorithm that was discussed in lecture 5. This is how the CNN learns to extract features that are most relevant to the task at hand.

Typically, CNNs include max pooling layers after each convolutional layer. And these max pooling layers serve exactly the same function that they did in HMAX: They help the network recognize the same features at slightly different spatial locations.

Hierarchy of CNN Layers

A CNN always arranges its convolutional layers and max pooling layers one after another in a very deep hierarchy of layers. Each subsequent set of layers in the hierarchy performs the same convolution and max pooling operations, but they work on the output of the preceding layers, not the image itself. That layered hierarchy means each layer can recognize more and more complex features that are built up from the lower-level features.

CNNs often contain dozens or hundreds of convolutional and max pooling layers. By the time you get to the top of the hierarchy, the network might be processing some very complicated features. But all these kernels are being learned, and so the features that get extracted are actually useful in distinguishing different visual stimuli, like a cat versus a tiger or a glass versus a bottle.

The last two layers in a convolutional neural network are often a fully connected layer and a softmax layer. In a fully connected layer, every neuron is connected to every neuron in the previous layer. The purpose of the fully connected layer is to allow the network to classify the input image. It therefore typically contains one neuron for every possible image classification. One neuron corresponds to "dog," one corresponds to "table," one corresponds to "guitar," and so on. Every possible image classification has a corresponding neuron in the fully connected layer.

And when the "dog" neuron is very active, that reflects the fact that the network strongly believes the image contains a dog. Conversely, if the "guitar" neuron is not active, that reflects the fact that the network doesn't believe there's a guitar in the image.

The softmax layer converts the activations over the fully connected layer into probabilities that add up to 1. So, if the network thinks the image could be a picture of a tree or a bush, then each of those neurons in the softmax layer should have a value near 0.5, while all the other neurons should have a value near 0.

Then, this entire deep network is trained on as many images as you can get your hands on—preferably millions of images. And each image needs to have a label indicating what the correct classification of the image is. Then, you can train the network and repeatedly change the synaptic weights according to the backpropagation rule. And with enough images and enough training, convolutional neural networks can get extremely good at visual object recognition.

Responses of Real Neurons versus Neurons in CNNs

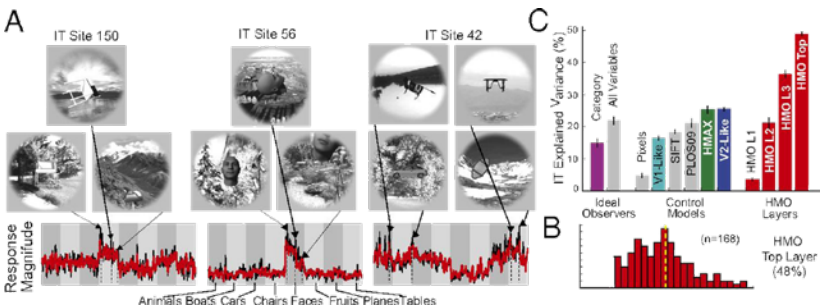
While convolutional neural networks that were inspired by studies of the brain's visual system are approaching humanlike object recognition abilities, it turns out that these networks are also shedding light on how vision is implemented in the brain.

In one famous study, Daniel Yamins, James DiCarlo, and their colleagues at the Massachusetts Institute of Technology optimized a CNN on a large dataset of natural images. They then compared the responses of the network to those of neurons in real brains. And amazingly, they found that artificial neurons in their CNN were able to replicate many of the response properties observed in real neurons. The similarity suggests that in addition to being useful as computer vision systems, convolutional neural networks may be a reasonably good model of how the brain solves the vision problem.

For example, a monkey was presented with a wide array of visual stimuli, including animals, boats, chairs, faces, fruits, and many other objects. The black traces on this graph plot the monkey's real neural responses from three different sites in inferior temporal cortex. The site on the left seems to respond a lot to chairs, the site in the middle seems to respond a lot to faces, and the site on the right responds to a variety of different visual stimuli.

Amazingly, the red traces on the graph correspond to the responses from three different artificial neurons in one of the highest layers in the CNN when it was presented with the same visual stimuli. Notice that responses of the artificial neurons match the responses of the real neurons almost exactly.

Remember, the CNN was never fit to the real neural data; it was just trained on visual images and optimized to accurately categorize what it saw. Yet, the responses of neurons in the CNN look just like the responses of real neurons!



Supervised versus Unsupervised Learning

One significant problem with CNNs is that, to perform well, they need to be trained on a very large set of visual images—often more than 10 million images. And every single image has to be labeled by a human being so that the CNN knows what the right answer is. That's how supervised learning algorithms like backpropagation work. You have to know the right answer to compare the network's output against ground truth. Then, you can compute the error and change the synaptic weights to reduce that error.

Unfortunately, most neuroscientists don't consider the backpropagation algorithm to be a biologically plausible mechanism. For one thing, the algorithm itself requires numerous complex calculations, like computing error gradients, and it seems like a stretch to assume that real neurons are performing those kinds of calculations.

An even bigger problem is the need for millions of category labels to serve as targets during training. That just isn't the way people naturally learn to recognize objects. After all, human babies get very good at recognizing their parents, pets, and objects in the environment long before their language skills are fully developed. And many nonhuman animals are extremely good at visual object recognition despite never learning any verbal labels at all.

So, is there any way that a deep convolutional neural network could learn to recognize objects without depending on an enormous database of labeled examples? And if so, would the responses of the artificial neurons still match the responses of real neurons like you saw for supervised CNNs? Amazingly, the answer to both questions appears to be yes.

But how can you possibly learn something if you don't know what the right answer is? It turns out that there are actually quite a few unsupervised learning methods that do just that, and most of these methods work by learning to represent the inputs more efficiently.

For example, suppose you want to represent a picture containing three squares. Imagine the input is a matrix of pixels that's 100 by 100 where the pixels in the squares have a value of 1, and all the other pixels have a value of 0. So, the input representation has 10,000 dimensions (100 × 100), one for each pixel. But you could represent those three squares in far fewer

dimensions if you just represent the *x*, *y* coordinate of the center of each square along with a number representing the size of the square, like its diameter. So, now, instead of 10,000 dimensions, you're representing the same information with only 9 dimensions, 3 for each square. And yet, with those nine numbers, you can perfectly reconstruct the original 100-by-100 picture.

Learning to represent input patterns efficiently means you use fewer dimensions, which is what unsupervised learning mechanisms do. So, the MIT group tried applying an unsupervised learning mechanism called contrastive embedding to a deep convolutional neural network. The goal of contrastive embedding is to learn a simple, low-dimensional representation of the inputs that can still distinguish between them.

To achieve simple, low-dimensional representations, you want to give very similar representations to different images of the same object. So, all images of a particular tree would have a very similar representation. Specifically, the researchers used an algorithm called local aggregation, which refers to the fact that the algorithm aggregates together the similar, low-dimensional representations of different views of the same object.

But you also want your representations of the inputs to be able to distinguish between different objects. So, contrastive embedding gives different objects very different representations. It's called contrastive because you're trying to make the low-dimensional representations of different objects as different as possible.

Once the network had learned representations that satisfied this contrastive property, they tested whether these representations could support object recognition and whether the learned representations looked similar to the representations in real brains.

And sure enough, the network could categorize input images quite accurately and perform other visual tasks, like determining the orientation or pose of an object, estimating the position of an object, and estimating an object's size.

The researchers also compared the responses of their artificial neurons to the responses of real neurons to the same set of visual stimuli. And, once again, they matched. In fact, the match with the unsupervised model was even better than the match they found using the supervised model.

The bottom line is that scientists have now addressed one of the major complaints about convolutional neural networks as a model of the brain's visual system. A supervised learning algorithm like backpropagation has questionable biological validity and requires millions of labeled inputs. But there are now CNNs that have been trained using much more plausible unsupervised methods. And these models also succeed in explaining both visual performance and the activity of neurons in the visual system of the brain.

Reading

Khan, S., H. Rahmani, S. A. A. Shah, M. Bennamoun, G. Medioni, and S. Dickinson. *A Guide to Convolutional Neural Networks for Computer Vision*. San Rafael, CA: Morgan & Claypool Publishers, 2018.

Marr, D. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Cambridge, MA: MIT Press, 2010.

Poggio, T. A., and F. Anselmi. *Visual Cortex and Deep Networks: Learning Invariant Representations*. Cambridge, MA: MIT Press, 2016.

17

What Damage Reveals about the Brain

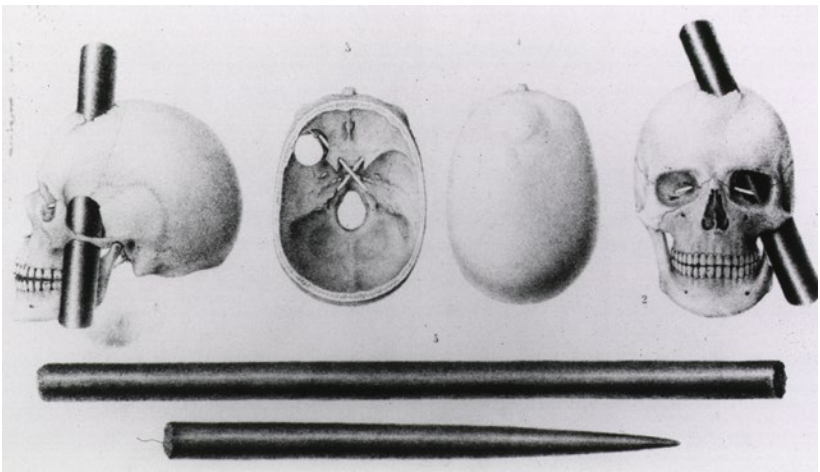
Throughout this course, you've explored how damage to particular cortical regions or pathways can produce various deficits in such areas as language processing, memory, and emotions. This lecture continues that discussion and focuses on some of the most fascinating neurological conditions that arise from strokes, tumors, and other types of brain damage. You'll learn about the behaviors associated with the conditions and what they tell scientists about the way the brain normally works.

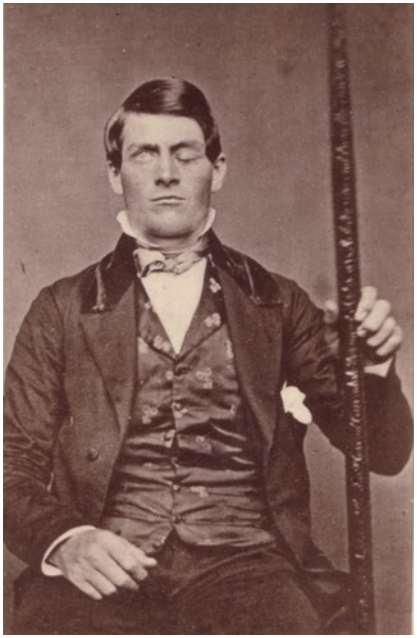
Phineas Gage

Phineas Gage was a railroad construction foreman who lived in the 19th century. He became famous after surviving one of the most shocking brain injuries you can imagine. On September 13, 1848, Gage was working on the construction of the Rutland and Burlington Railroad near Cavendish, Vermont. He was blasting rock to make way for the railway, which involved boring a hole into the rock, adding blasting powder, and then packing in some kind of inert material into the hole using a tamping iron.

The tamping iron that Gage was using was 3 feet, 7 inches long and a little more than 1 inch in diameter. It weighed more than 13 pounds. As he was tamping down the material in one of these holes, the blasting powder exploded and sent the tamping iron rocketing out of the hole, up through the left side of Gage's face, behind his left eye, through the left frontal lobe of his brain, and out the top of his head. It landed about 80 feet away.

Amazingly, Gage survived. He even remained conscious and was able to speak soon after the accident. He was taken to a nearby physician, Dr. John Martyn Harlow, who provided initial medical care. Gage's wound was cleaned and dressed, but he was left with permanent damage in the frontal lobe of his brain.





Harlow reported significant changes in Gage's personality after the accident. Whereas he said that before the accident Gage was responsible, hardworking, and an excellent railroad foreman, he described Gage after the accident as "fitful, irreverent, indulging at times in the grossest profanity" and like "a child in his intellectual capacity and manifestations." He also noted that Gage's "friends and acquaintances said he was 'no longer Gage.'"

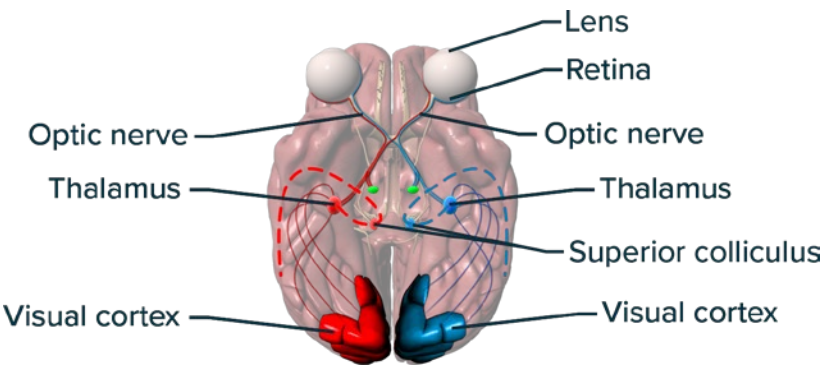
Harlow's description of Gage supports the hypothesis that the frontal lobes play a major role in a person's personality. And substantial evidence now supports that claim. It seems likely that he did suffer from some significant personality changes that made him hard to get along with at first, but apparently these problems either resolved or became less severe over time. And despite one of the most horrific brain injuries one can imagine, Phineas Gage went on to live a relatively normal life.

Four years after the accident, Gage moved to Chile and worked as a long-distance stagecoach driver. For the next 7 years, he continued in this job, which required a great deal of responsibility. His case underscores the idea that damage to certain parts of the brain can lead to some surprising and unusual phenomena that can teach scientists important lessons about the normal brain.

Blindsight

Patients with a condition called blindsight demonstrate characteristics of sight despite being essentially blind. To understand what's going on, you need to first understand how vision is normally implemented in the brain.

When light enters the eyes, it's focused by the lens and falls on the retina, a thin layer of cells at the back of the eye. The retina contains photoreceptor cells—the rods and cones—which convert the light into electrical signals that are sent to the brain through the optic nerve. These signals typically first go to the thalamus, which is like the brain's switchboard operator. After passing through the thalamus, most visual signals are sent to the primary visual cortex at the back of the brain in the occipital lobe.



The visual cortex is responsible for analyzing the signals and creating a visual perception of the world around you. Damage to it can lead to significant problems with vision. For example, a stroke that destroys the visual cortex in the left hemisphere might make a patient blind to the right side of space, and the inverse would be true for the right visual cortex and the left side of space. This is called a hemianopia, meaning blindness in half the visual field.

One patient, who goes by the initials D. B., had to have brain surgery to remove a vascular malformation in his right occipital lobe. Afterward, he had a visual exam to determine his visual field. Each eye was covered one at a time, and he was asked to keep his open eye fixed on a single point without moving it. Then, the examiner slowly moved a visual stimulus from outside the visual field toward the center of the visual field, and D. B. had to report when he first saw the stimulus. The researchers created a reliable map of the parts of the visual field that the patient could and couldn't see with each eye.

D. B.'s visual exam revealed a homonymous hemianopia in his left visual field, meaning that he didn't report seeing stimuli in the left visual field of either eye. And that's what you might expect after surgical damage to the right occipital lobe.

But Larry Weiskrantz, a professor at Oxford University, asked D. B. to guess about visual stimuli in his left visual field. And despite having no conscious visual experience, D. B. often guessed correctly and performed significantly above chance. D. B. had blindsight—he demonstrated some kind of vision in his damaged field.

D. B.'s case and others have taught scientists some important lessons about vision and how it's implemented in the brain. First of all, the cases demonstrate that conscious awareness is not a prerequisite for some types of visual processing. Despite not being consciously aware of the stimuli in their impaired visual field, blindsight patients can still respond to them in some ways.

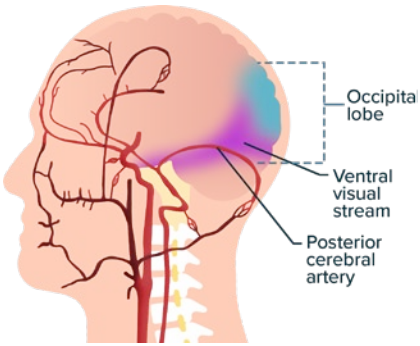
A second lesson is that some of the visual circuits in the brain must bypass the primary visual cortex in the occipital lobe. Some evidence suggests that subcortical visual pathways that include a brain structure called the superior colliculus in the midbrain probably play a role in some of the visual abilities exhibited by blindsight patients. Analogous phenomena have also been reported in other sensory modalities, including touch and hearing.

In patients with blindsight, visual stimuli get processed to some extent, but that processing happens below the level of consciousness.

Visual Agnosia

John was a patient who was studied by Glyn Humphreys and Jane Riddoch at London University. John served as a pilot for the British air force during the Second World War and then worked for a company that sold windows for houses. He later rose to an executive position with an American company and was responsible for their marketing in Europe.

John's life changed dramatically in April 1981, when he had an appendectomy and suffered a stroke involving his posterior cerebral artery. The stroke damaged the so-called ventral visual cortex but spared more dorsal parts of the occipital lobe.



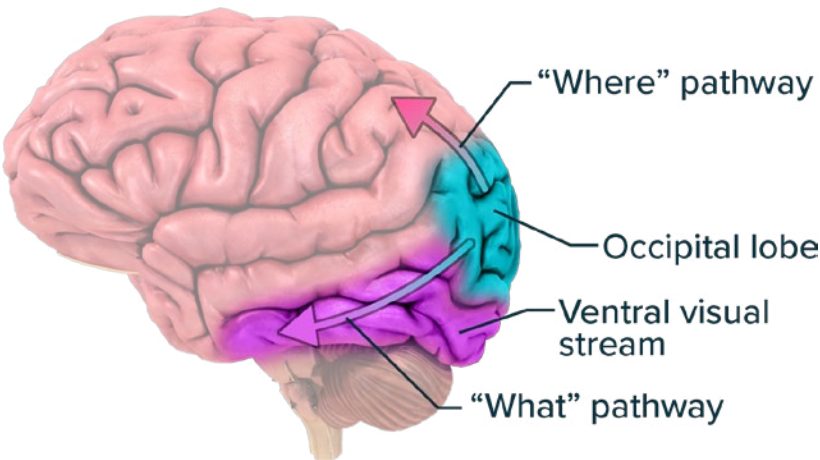
At first, John was completely unaware that he'd had a stroke, and even his doctors didn't realize it for some time. And that's because he didn't have many of the typical symptoms, such as facial drooping, muscle weakness, or speech difficulties. Furthermore, his memory and intelligence seemed relatively unaffected.

But John's wife, Iris, noticed something was wrong. John couldn't recognize objects by sight—not even very familiar objects, like flowers and letters and words. Visual agnosia also impedes one's ability to recognize faces, and John was unable to recognize the faces of his doctors and nurses. Worse yet, he no longer recognized his wife by sight. Even when he looked at himself in the mirror, his own face seemed unfamiliar.

Most patients with visual agnosia can accurately describe what they're looking at and can even draw a good copy. But they have a hard time recognizing what it is that they're looking at. Their spatial vision is also typically spared. That is, they can still tell you where things are even though they can no longer tell you what they are. They can also recognize objects by hearing or by touch. So, for example, even though John couldn't recognize his wife by sight, as soon as she spoke a word, he immediately knew who she was.

Some patients have a deficit that is restricted exclusively to faces. This kind of problem is referred to as prosopagnosia, from the Greek words for "face" and "ignorance." Conversely, strokes can sometimes lead to a visual impairment that is restricted to letters and words, which is known as an alexia. In some cases, the person might have severe problems with reading even though their spoken language skills are completely intact and even though they can still write. Although, when they do write, they typically can't read what they just wrote.

These cases of visual agnosia tell scientists a number of things about the normal operation of the brain. First, they suggest that perception and recognition can be dissociated. Second, visual agnosia also suggests that humans process different types of visual information using different neural circuits. For example, the ventral visual stream that is damaged in agnosia seems to be particularly important in figuring out what things are in a person's visual environment. In fact, the ventral stream is sometimes referred to as the "what" pathway for that reason. But even when that pathway is damaged, people can still figure out where things are. So, spatial vision is different than object vision and uses a different neural circuit.



And even within the ventral visual stream, it appears that different neural circuits are used for processing different types of visual stimuli. In particular, some patients lose the ability to recognize faces despite the fact that they can still read. Conversely, other patients lose the ability to read despite still being able to recognize faces. So, apparently recognizing words and recognizing faces depend on different neural circuits.

Hemispatial Neglect

Dr. Jenni Ogden, a neuropsychologist, studied a patient named Janet and described her in the book *Fractured Minds*. On Janet's 50th birthday, she drove home from work, and as she was parking her car, she collided with the left side of her garage. Later that evening, as she tried to blow out the candles on her birthday cake, she initially blew out only the candles on the right.

A couple of days later, her husband found her lying on the bathroom floor after apparently having a seizure. He drove her to the hospital, and a CT scan revealed a large tumor in her right parietal lobe. Neurosurgery was performed, which removed most of the tumor, but, unfortunately, it grew back. And it had numerous effects on her behavior.

For example, Ogden asked Janet to read out loud from a book. And although the words she read were correct, she typically failed to read the first two or three words on every line. Likewise, when writing, Janet tended to write exclusively on the right side of the paper and would leave the left side blank.

Janet seemed to ignore, or neglect, the left side of space. And consistent with this pattern, Janet's condition is often referred to as hemispatial or unilateral neglect. Patients with this condition tend to ignore many or even most of the stimuli on the left side of space.

Patients with hemispatial neglect will often eat food only on the right side of their plate. In some cases, when dressing, they might not put their left arm in the sleeve of their shirt or their left leg in their pants.

Now, you might suppose these patients behave this way because they are simply blind in their left visual field. But there are plenty of patients with left-sided blindness who don't behave this way. They compensate for their partial blindness by turning their head and bringing whatever is to their left into their good visual field. But patients with hemispatial neglect don't do that; they simply ignore the information on the left.

Furthermore, neglect patients don't always ignore the information on the left. So, one day they'll fail to copy the left side of a picture, but the next day, they manage to do so. This kind of variability suggests that the problem is not perceptual. Rather, it's a problem of attention. They ignore information that's on the left even though they're capable of processing it.

Janet's case demonstrates that humans are not always consciously aware of what they perceive. Of course, everyone regularly fails to notice many aspects of the world even when they perceive them, but the neglect syndrome really highlights the importance of attention in conscious experience.

Neglect also reveals something important about the way attention is implemented in the brain. Specifically, it suggests that the right parietal lobe plays a particularly important role in allowing people to pay attention to the left side of space, another example of the brain's contralateral organization.

Reading

Macmillan, M. *An Odd Kind of Fame: Stories of Phineas Gage*. Cambridge, MA: MIT Press, 2002.

Ogden, J. A. *Fractured Minds: A Case-Study Approach to Clinical Neuropsychology*. New York: Oxford University Press, 2005.

Sacks, O. *The Man Who Mistook His Wife for a Hat and Other Clinical Tales*. New York: Summit Books, 1985.

Weiskrantz, L. *Blindsight: A Case Study Spanning 35 Years and New Developments*. Oxford: Oxford University Press, 2009.

18

Depression and Anxiety

This lecture introduces you to computational psychiatry—an interdisciplinary subfield of cognitive science that combines methods and concepts from computer science, mathematics, and neuroscience. The hope is that computational and mathematical models will help researchers understand the mechanisms that underlie psychiatric disorders, such as depression, anxiety, and schizophrenia so that new treatments can be developed to help the millions of people suffering from mental illness.

Major Depressive Disorder

Everyone feels sad at times. It's a perfectly normal, and probably healthy, reaction to life's disappointments. But in people with major depressive disorder, feelings of sadness, hopelessness, or emptiness persist for long periods of time and even when circumstances change. These feelings are so strong that they interfere with the person's ability to participate in their normal daily activities. In this lecture, the term *depression* is used as a shorthand for major depressive disorder, but it's important to remember that the clinical disorder is fundamentally different from the normal sadness that everyone feels from time to time.

Two other characteristic features of depression are anergia, or a lack of energy or vitality, and psychomotor retardation, a slowing down of physical and mental processes. Patients also often exhibit negative biases in their thinking and in the way they view the world. For example, they tend to pay more attention to negative facial expressions and words than to those that are positive. They are also more likely to remember negative events and words compared with people without depression.

This kind of negative bias also clouds their view of the future. Patients with depression tend to think that negative events are more prevalent than positive events and that they are more likely to happen in the future.

Another common symptom of depression is rumination—repetitive thinking that involves dwelling on negative thoughts, emotions, and experiences without finding a solution or a way to move past them. This type of thinking can lead to persistent feelings of hopelessness, helplessness, and worthlessness.

People with depression are more likely to interpret ambiguous information, such as a neutral facial expression or remark, as negative or critical even if it wasn't meant that way.

Treatment for depression often involves a combination of medication, therapy, and lifestyle changes. Unfortunately, those treatments don't work for everyone, and scientists don't really know why. Part of the problem is that researchers don't yet have a clear understanding of the mechanisms that underlie the disease, which makes it challenging to develop effective treatments. But cognitive scientists in the field of computational psychiatry are hoping to change that by developing computational models of the disorder.

Applying Bayesian Decision Theory to Depression

One popular computational model for depression was proposed by Quentin Huys, Nathaniel Daw, and Peter Dayan and is based on a framework called Bayesian decision theory. This theory combines two different ideas that were covered in the discussion about decision-making in lecture 13: reinforcement learning and Bayesian analysis.

The basic idea behind Bayesian decision theory is that people combine their beliefs about the current state of the world with their prior beliefs in order to estimate the true state of the world. Then, they choose actions that maximize long-term reward given their beliefs about the state of the world. And sometimes they can improve their choices with model-based reasoning, in which they simulate how the world might change based on the choices that they make.

Huys, Daw, and Dayan argued that many of the major symptoms of depression would naturally follow from Bayesian decision theory if you assume maladaptive prior beliefs that overall rates of utility are generally low. That is, the person with depression believes that the long-term value or utility of almost any action won't be very high. When you combine this assumption with the other basic elements of Bayesian decision theory, it provides an interesting explanation for many of the major phenomena associated with depression.

First, consider the primary symptom of persistent feelings of sadness, hopelessness, or emptiness. That maps directly onto the belief that the long-term value of virtually any action is low. After all, if you believe that no matter what you do or what decisions you make, nothing is going to make much difference or significantly improve your life, then it seems appropriate to feel sad, hopeless, and empty.

Furthermore, because Bayesian computations always consider the priors when computing value estimates, those value estimates will tend to stay low, even when circumstances change. So, even if the current state of the world changes and the depressed person is offered an opportunity that could positively impact their life, their estimate of the value of that opportunity will be dragged down by their prior belief that no choice could have a substantially positive impact. As a result, they might not seize an opportunity that could be worthwhile.

That kind of pessimistic prior belief will also impact model-based reasoning about the future. In particular, it could have a significant impact on what the depressed person thinks is likely to happen. For example, suppose you're trying to simulate how the world might change based on different choices in order to figure out each choice's utility. It's obviously crucial that you accurately estimate the utilities associated with those future states. But if your prior assumption is that all future states are relatively unrewarding, then model-based reasoning will continue to assign negative values to all the different actions you consider, and you won't be able to get out of the negative loop.

Assigning low values to all possible actions can also lead to rumination over negative events. Now, everyone occasionally thinks about past negative events and has regrets about some past choices, but most people also recognize that those kinds of thoughts usually don't have a lot of utility, and so they typically inhibit them and explore alternative thoughts when they're doing model-based reasoning.

But for people with depression, the utility of those ruminative thoughts is no more negative than the utility of other thoughts that people without depression would consider more productive, and so depressed people might run that negative event over and over in their mind.

It also turns out that maladaptive, pessimistic priors can provide insights into anergia and psychomotor retardation. In particular, people are typically most motivated to pursue a course of action when they believe that the course of action will be rewarding. Conversely, if a depressed person implicitly believes that no course of action will lead to significant reward, then they would likely exhibit significantly reduced motivation. And such a lack of motivation might manifest itself as a lack of energy or vitality. Likewise, it might lead to less vigorous responding, which could show up as slower motor movements.

Unfortunately, according to the theory, this lack of motivation and associated lethargy could lead to a kind of vicious cycle. To update your estimates of the long-term value associated with different actions, you have to try out a variety of actions and learn about the different utilities associated with each. But if you're not motivated to pursue any action, and if you tend to take actions more slowly, then you're also going to learn more slowly. And as a result, it's going to be hard to update your maladaptive priors, which assume that most choices have low utility.

Anxiety Disorders

It's important to distinguish anxiety from fear. Clinicians typically use the term *fear* to refer to an emotional response to a recognized external threat. For example, you might experience fear if you see a rattlesnake on the ground nearby or if someone points a weapon at you. In those cases, there's a clear external threat, and being afraid is appropriate; in fact, in those cases, fear is an adaptive response that could potentially help you deal with the threat.

Anxiety, on the other hand, refers to an unpleasant emotional state that's similar to fear but for which there is no easily identifiable cause. And in anxiety disorders, the feelings of anxiety are strong and pervasive and significantly interfere with the activities of daily life. Patients with anxiety disorders are often treated using psychotherapy, medication, or a combination of both.

Anxiety disorders are associated with a number of other symptoms, including restlessness, fatigue, irritability, difficulty concentrating, and sleep disturbances. Anxiety disorders are also associated with maladaptive choice behavior, and that maladaptive decision-making might provide insight into the underlying problem.

One of the main ways pathological anxiety affects decision-making is illustrated by so-called avoidance behaviors. Individuals with anxiety tend to avoid making decisions or taking risks that have the possibility of leading to negative outcomes, even if those outcomes are unlikely. For example, someone with an anxiety disorder might avoid social situations, such as parties or meetings, because they're worried about feeling embarrassed or rejected. Or someone might avoid flying, even though the chance of an accident is extremely remote.

Anxiety also dramatically impacts the way people simulate future events during model-based reasoning. In particular, people with anxiety tend to put inappropriate emphasis on negative outcomes that could arise in the future and have a hard time imagining potential positive outcomes when making decisions. They also tend to spend an inordinate amount of time worrying about these potential future negative outcomes. This worry can lead to decision paralysis, which can occur when individuals become so overwhelmed by their anxiety that they struggle to make any decision at all.

Applying Bayesian Decision Theory to Anxiety Disorders

Developing computationally explicit models of anxiety disorders is still relatively new, but progress is being made. One approach that has had some success was proposed by Sonia Bishop and Christopher Gagne at the University of California, Berkeley. Their model is based on the same kind of Bayesian decision theory that was discussed for depression.

But whereas depression is associated with universally low estimates of the utility of different actions, anxiety is associated with increased estimates of the future probability of negative outcomes as well as biased estimates of how negative those outcomes will be.

As a result, people suffering from anxiety disorders tend to avoid choices and actions that other people would feel comfortable with. And repeated avoidance behavior reduces opportunities to update their estimates of both the value and probability of negative outcomes. For example, if you never go to parties, then you'll never have the opportunity to learn that your predictions about them were wrong. As a result, you'll never update your estimates of either the probability or the utility of the negative outcomes. So, when the next party rolls around, nothing has changed, and you'll tend to engage in avoidance behavior again.

This theory also provides a natural explanation for why individuals with anxiety have difficulty imagining potential positive outcomes. In particular, during model-based reasoning, positive outcomes are viewed as much less likely than negative outcomes, and so they get pruned out when someone is thinking about what is likely to happen in the future. Unfortunately, this tendency only reinforces the avoidance behavior. And as a result, it increases the likelihood of missing opportunities for correcting the maladaptive estimates.

Reading

Berger, J. O. *Statistical Decision Theory and Bayesian Analysis*. New York: Springer Science & Business Media, 2013.

Martin, C. R., L. A. Hunter, V. B. Patel, V. R. Preedy, and R. Rajendram, eds. *The Neuroscience of Depression: Features, Diagnosis, and Treatment*. London: Academic Press, 2021.

Redish, A. D., and J. A. Gordon, eds. *Computational Psychiatry: New Perspectives on Mental Illness*. Vol. 20 of *Strüngmann Forum Reports*. Cambridge, MA: MIT Press, 2016.

Seriès, P. *Computational Psychiatry: A Primer*. Cambridge, MA: MIT Press, 2020.

19

Autism, Schizophrenia, and OCD

This lecture continues the discussion of mental disorders by focusing on three common psychiatric conditions: obsessive-compulsive disorder, schizophrenia, and autism. You'll learn about the symptoms of each disorder as well as what cognitive scientists think might be happening in the brain to cause the experiences and behaviors associated with each condition.

Obsessive-Compulsive Disorder

Obsessive-compulsive disorder (OCD) is a mental health disorder characterized by obsessions (which are persistent, unwanted, and intrusive thoughts) and compulsions (which are repetitive behaviors or mental acts). The obsessive thoughts often involve concerns about contamination. While some patients with OCD fear germs and illness, others have obsessive thoughts about being hurt or injured. Another common obsession is with symmetry and order. Patients can become obsessed with having everything on their desk or night table in exactly the right place and organized in the same way every day.

People with OCD also typically perform repetitive behaviors, or compulsions, that are aimed at reducing the distress or anxiety caused by their obsessions. For example, patients who are obsessed with germs might repeatedly wash their hands dozens of times a day.

Another common compulsion is checking behavior. For example, patients often feel compelled to check their locks over and over to make sure that they're locked. Or they might repeatedly check that they turned off their stove or their lights.

These compulsions can often turn into rituals that people with OCD feel compelled to perform in exactly the same way and often a specific number of times. And if any of the ritual is performed incorrectly or just doesn't feel right, then they might feel compelled to do it all over again until they get it exactly right.

The victims of OCD typically realize intellectually that their obsessive thoughts and compulsive behaviors don't make logical sense, but they just can't help themselves. The feelings of distress and anxiety that they experience are very real, and they are desperate to try to alleviate them in any way they can, even if it doesn't make any logical sense. The disease can make life absolutely miserable for the victim.

OCD affects about 1% to 2% of the population worldwide and usually emerges during childhood or young adulthood.

Computational Models of OCD

Cognitive scientists don't yet know for sure what exactly is going on in OCD, but they are beginning to develop computationally explicit models of the disorder that could unravel the mystery. One important model was proposed by Isaac Fradkin and Jonathan Huppert at the Hebrew University of Jerusalem along with a number of collaborators at the University College London. They conceptualized OCD in a Bayesian framework in which people are never completely certain about the current state of the world or about the outcome of the actions that they take. Instead, they compute probabilistic estimates of what the current state is by combining evidence from their senses with their prior knowledge and expectations.

Using Bayes's theorem, the researchers combine current evidence from the senses with prior base-rate knowledge to develop better estimates about what the state of the world really is. One critical assumption that has been added is that there's more weight put on more certain information and less weight put on less certain information. For example, suppose you're trying to make your way through a dark room to the bathroom in the middle of the night. If you're in your own bed at home, and you know exactly where the bathroom is, then you can rely a lot on memory and may not even need to turn on the light.

But if you're sleeping in an unfamiliar room and are less certain about where the bathroom is, then you'll probably need to turn on the light and look around to get your bearings. The bottom line is that, when you're fairly certain about the state of the world, you don't need to rely as much on sensory input as you would when you're much less certain.

Fradkin, Huppert, and their collaborators proposed that people with OCD have excessive uncertainty about how the state of the world changes after different actions occur, particularly their own actions. And because they're much less certain about these transitions than other people, they rely much more on sensory input than other people do.

They're like the person sleeping in an unfamiliar room. Because they tend to be very uncertain about where the bathroom is, they need to rely much more on sensory input. And this transition uncertainty underlies the major symptoms associated with OCD.

Fradkin and colleagues illustrate the problem with an example. Suppose you're leaving home, and you put your wallet in your bag. For most people, once you put your wallet in your bag, you feel quite certain about the new state of the world with respect to your wallet: It's in your bag. And just like you don't need to turn on the light to find the bathroom at home, you don't need to check to make sure your wallet is in the bag. In terms of the Bayesian theory, you have low transition uncertainty, and so you don't feel the need to gather additional sensory information.

But now consider the situation for a person with OCD who is very uncertain about state transitions. They might remember putting their wallet in their bag, but they still feel uncertain about whether the wallet is actually there. And in the absence of concrete sensory evidence demonstrating that the wallet is still in the bag, they feel compelled to check to make sure that it is. In other words, they depend much more on direct sensory information than other people do, because they're much less certain about the outcomes of previous actions. And checking gives them direct sensory confirmation.

Schizophrenia

Schizophrenia is a severe, debilitating psychiatric disease characterized by a range of symptoms that affect a person's thoughts, emotions, and behaviors. It's associated with three major categories of symptoms: positive, negative, and cognitive.

Positive symptoms include abnormal experiences that are added to a person's reality. They are considered positive not because they're good but because they go beyond what most people experience. For example, patients with schizophrenia often experience delusions and hallucinations. Delusions are false beliefs that the patient firmly holds despite evidence to the contrary. For example, they might believe that they are being followed or that someone's trying to harm them. Hallucinations might include hearing voices or even seeing things that are not actually there. Most people without schizophrenia don't experience these things.

In the case of negative symptoms, people with schizophrenia experience less than a person without the disease: less pleasure, less motivation, and less emotion. The five main types of negative symptoms in schizophrenia are referred to as the five As: affective flattening, alogia, anhedonia, asociality, and avolition.

Affective flattening refers to a lack of emotional expression. When most people would express strong emotions, like joy, anger, or fear, people with schizophrenia will often exhibit a significantly blunted or flattened response.

Alogia refers to a reduction in the amount or content of speech. This is also sometimes referred to as poverty of speech. For example, someone with schizophrenia might talk less frequently, be less spontaneous in conversation, or limit themselves to brief one-word responses to questions.

Anhedonia refers to a lack of pleasure or interest in activities that are normally enjoyable. Patients may find it difficult to experience positive emotions and may have a reduced ability to anticipate pleasure from future events.

Asociality refers to a lack of interest in social relationships or a reduced desire for social interaction. Patients may prefer to be alone or may have difficulty initiating and maintaining social relationships. They may not feel a need for social interaction or may not see the benefits of engaging in social activities.

Finally, avolition refers to a lack of motivation or an inability to initiate and persist in goal-directed activities. Patients may have difficulty completing tasks or may not feel a sense of satisfaction or accomplishment from completing them. They might exhibit a lack of interest in work or school or a lack of concern with personal hygiene.

Cognitive symptoms include difficulties with thinking, learning, memory, and decision-making. Patients with schizophrenia often exhibit slow or disorganized thinking. They often have poor memory and a hard time concentrating. They also have difficulties with communication, both in understanding what others say and in expressing their own thoughts. Many of these problems are related to problems with executive control more generally. These cognitive symptoms can be particularly disabling because they can interfere with a person's ability to function independently and to perform the basic activities of daily life.

The exact causes of schizophrenia are not well understood, but, like most psychiatric disorders, both genetics and environment play a role.

Models for Understanding Schizophrenia

Two common models attribute schizophrenia to changes in two of the brain's major neurotransmitter systems: dopamine and glutamate. In the 1950s scientists accidentally discovered that a type of drug called a phenothiazine, which was originally used as an antiseptic and to treat malaria, also happened to help treat the positive symptoms of schizophrenia. It was later discovered that these drugs block receptors that respond to dopamine.

Around the same time, the Swedish neuropharmacologist Arvid Carlsson discovered that amphetamine, which was known to produce schizophrenia-like hallucinations and delusions in chronic users, triggered the release of dopamine. Together, these independent lines of research provided strong evidence that excess release of dopamine might play an important role in schizophrenia.

One weakness of the dopamine model, however, is that it only really addresses the positive symptoms of the disease. So, what explains the negative and cognitive symptoms of schizophrenia? Some insight into that question came from studies of the drug PCP, better known as angel dust, as well as the anesthetic ketamine, which has effects similar to those of PCP.

Scientists found that both drugs can produce the full range of symptoms associated with schizophrenia, not just the positive symptoms. For example, people taking these drugs often grow withdrawn, antisocial, and apathetic, mimicking some of the negative symptoms of schizophrenia. They also exhibit deficits in abstract thinking, in learning, and in memory, mimicking some of the cognitive symptoms.

Both PCP and ketamine significantly impact the brain's glutamate system. Specifically, they block one type of glutamate receptor that's known as the NMDA receptor. And as a result of these findings, many scientists now believe that the negative and cognitive symptoms of schizophrenia are primarily due to a disruption in the brain's glutamate system.

But it turns out that NMDA receptors also play a role in regulating the release of dopamine. And blocking NMDA receptors can produce some of the positive symptoms associated with schizophrenia. The bottom line is that disruption in the glutamate system, and specifically in NMDA receptors, provides a relatively simple explanation for the full range of symptoms associated with the disease.

Autism Spectrum Disorder

Autism is a developmental disorder that affects communication, social interaction, and behavior. It's referred to as a spectrum disorder because it presents in a wide spectrum of different ways in different individuals. Nevertheless, a few symptoms are quite common, and they play an important role in communication.

First, most people with autism experience some kind of difficulty with social interaction. They might have trouble making eye contact, understanding social cues that most people instinctively pick up on, and engaging in a reciprocal conversation. In extreme cases, autistic people may have difficulty interacting with other people at all and may appear to be off in their own world.

People with autism also commonly exhibit repetitive behaviors, such as hand-flapping, rocking back and forth, or repeating words or phrases over and over. These kinds of behaviors are sometimes called stimming, or self-stimulatory behaviors. Many autistic people also develop a profound interest in specific topics or objects, which are sometimes referred to as fixations.

One of the most amazing facts about autism is that roughly 10% of autistic people also have some unusual ability or talent. This is referred to as savant syndrome. Although savant syndrome does sometimes appear in people with developmental disabilities or acquired brain damage, it's most common in people with autism.

Savants might exhibit an amazing memory for historical dates, sports trivia, or maps. A number of savants possess superhuman calendar skills, such as the ability to tell you the day of the week that a particular date falls on.

Cognitive scientists are working hard to develop computationally explicit models of autism, but it's fair to say that they still don't have a very good understanding of exactly what's going on. However, autism is not caused by vaccines. There is no credible scientific evidence for a link between the MMR vaccine and autism. Unfortunately, millions of people were led astray by the fraudulent claims of a now discredited doctor named Andrew Wakefield, and they decided not to vaccinate their children. And reduced vaccination rates probably contributed to significant outbreaks of measles and many unnecessary deaths.

Reading

Baron-Cohen, S. *Mindblindness: An Essay on Autism and Theory of Mind*. Cambridge, MA: MIT Press, 1997.

Javitt, D. C., and J. T. Coyle. "Decoding Schizophrenia." *Scientific American* 290, no. 1 (2004): 48–55.

Rapoport, J. L. *The Boy Who Couldn't Stop Washing: The Experience and Treatment of Obsessive-Compulsive Disorder*. New York: Penguin Group, 1991.

Treffert, D. A. *Islands of Genius: The Bountiful Mind of the Autistic, Acquired, and Sudden Savant*. London: Jessica Kingsley Publishers, 2010.

20

The Puzzle of Consciousness

This lecture dives into two of the most profound questions in all of cognitive science: What is the nature of consciousness, and how is it implemented in the human brain? To answer these questions, some studies investigate the contents of consciousness—that is, what people or even animals can report about what they are experiencing at a given moment. Other studies investigate how a person's level of consciousness affects brain activity in the hopes of learning something important about the neural correlates of consciousness.

Studying the Contents of Consciousness

Patients with severe and intractable epilepsy occasionally undergo a procedure called a corpus callosotomy. This split-brain surgery severs the corpus callosum, isolating the left and right brain hemispheres from each other. This surgery can prevent seizure activity from spreading from one side of the brain to the other.

But split-brain surgery also provides a unique opportunity to see into the amazing relationship between the mind and the brain. And it can potentially tell scientists something fundamental about the nature of consciousness. For example, while most people report experiencing a single, unitary consciousness, split-brain patients appear to experience two. Does a split brain also imply a split mind? The answer appears to be yes.

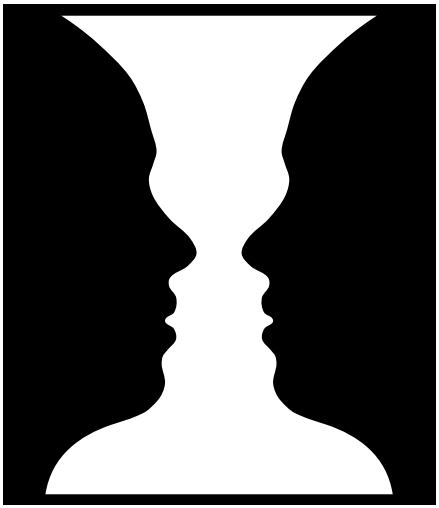
For example, imagine briefly flashing a picture to the patient's left and asking them to name it. In all brains, split or not, visual information from the left side of space gets processed by the brain's right hemisphere. But language is typically controlled by the left hemisphere. So, before the flashed object can be named, information needs to be communicated from the right hemisphere to the left. But without a pathway between the two hemispheres, split-brain patients typically won't be able to name objects presented in their left visual field. In fact, the speaking left hemisphere will even claim that it didn't see anything.

Amazingly, these same patients can still draw a picture of the object with their left hand, which is controlled by the right hemisphere. Their right hemisphere is consciously aware of the object even though the left hemisphere isn't. Apparently, these patients do not have a single, unitary consciousness. They have at least two.

Studies of neural activity in animals have taken this kind of work a step farther by identifying specific neurons whose activity is associated with the contents of consciousness. Nikos Logothetis and Jeffrey Schall at the Massachusetts Institute of Technology conducted an influential study using a binocular rivalry paradigm.

Binocular rivalry occurs when the left and right eyes receive different visual information, and so the brain has to resolve the rivalry between the eyes to interpret what is being seen. Think of looking through an old-fashioned stereoscope where each eye sees a slightly different version of the same image, leading to a perception of depth. However, in a binocular rivalry paradigm, the two eyes see different images that cannot be integrated.

You've probably seen the famous picture that can be seen either as two faces looking at each other or as a vase. Or perhaps you've seen the figure that can be interpreted either as a young woman or as an older woman. You can perceive the image either way, but you cannot perceive it both ways simultaneously. Rather, your perceptual experience switches back and forth.



Visual input stays exactly the same, but your conscious awareness of the input changes back and forth. It's therefore possible to identify neural activity that can be unambiguously associated with conscious awareness and that can't be attributed to the visual input itself.

Logothetis and Schall presented monkeys with sets of horizontal bars that were moving either up or down, and they recorded the activity of neurons in the visual system that were sensitive to motion. They trained the monkeys to move their eyes to the right if the bars were moving up and to move their eyes to the left if the bars were moving down.

Next, they put stereoscopic goggles on the monkeys. In each eye, the monkeys would see a different visual input. And just like human beings, the monkeys' perceptual experience switched back and forth. Sometimes, they perceived upward motion, and other times, they perceived downward motion. The visual input wasn't changing, only their perceptual experience was.

Interestingly, when the researchers looked at the activity of the motion-sensitive neurons, they found that some of them were active only when the monkey reported seeing upward motion and others were active only when he reported downward motion. Logothetis and Schall had found neurons whose activity was associated with perceptual experience—that is, with the contents of their conscious awareness!

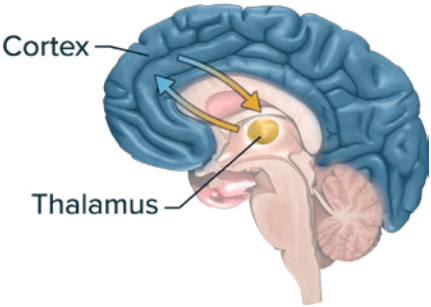
Studying Levels of Consciousness

Studies that investigate the level of consciousness are looking at what is happening in the brain when people are awake and conscious versus when they are unconscious, perhaps because they are asleep or anesthetized.

A number of studies have investigated this question using neuroimaging techniques like positron-emission tomography (PET) and functional magnetic resonance imaging (fMRI). The studies found that activity in a particular brain structure called the thalamus seems to be critically important in determining people's level of consciousness. When they're awake and conscious, there is substantial activity in the thalamus. When they are anesthetized or asleep, there is much less thalamic activity.

The thalamus is an egg-shaped structure almost exactly in the middle of your brain, and it functions like a kind of relay station, transmitting information to your cerebral cortex, which is the outer layer of your brain. This is where thinking happens, language gets processed, what you see and hear gets recognized, decisions get made, and voluntary movements are

programmed. It's the seat of both your intelligence and your personality; it's what makes you you.



The cerebral cortex is also where most conscious processing takes place. It's what got disconnected in the split-brain patients. It's where animal studies found neurons associated with perceptual experience.

The thalamus sends information from the outside world to the cerebral cortex. So, if anesthesia or sleep turns off the thalamus, that's going to cut off the flow of information and essentially isolate the cerebral cortex from the outside world. And that's exactly what happens when you're unconscious.

Scientists still have a long way to go before they have a complete understanding of the neural mechanisms underlying conscious experience. But these examples illustrate that they are making substantial progress.

Capturing the Subjective Experience

Imagine for a moment that neuroscientists have mapped out the neural correlates of conscious experience in great detail and that they can even manipulate someone's conscious experience. Will science then have "solved" the mystery of consciousness? Many neuroscientists and philosophers say no.

For example, the philosopher David Chalmers argues that science still hasn't scratched the surface of one of the most fundamental questions about consciousness: How does physical matter (the brain) give rise to subjective experience? Chalmers refers to this question as the hard problem of consciousness. Maybe scientists can find the answers to the "easy" problems by studying the neural correlates that go along with one subjective experience or another. But they still wouldn't understand why or how those neural correlates give rise to that subjective experience.

The philosopher Thomas Nagel wrote a very influential paper titled, "What Is It Like to Be a Bat?" He argues that, even if you know all the objective information there is to know about the brains and behavior of bats, that still doesn't tell you what it's like to be that bat.

This kind of conscious, subjective experience is what philosophers refer to as qualia, which comes from a Latin word meaning "of what kind" or "of what sort." It's the "what it's like" part of the experience. For example, what is it like to feel sad or happy, to see the color red, or to taste an apple? And is your experience of the color blue the same as that of your friend? What if your experience of blue is actually like their experience of red, but you've switched the labels? Would there be any way for science to tell the difference?

Many thought experiments have illustrated the difficulty of developing scientifically rigorous explanations of subjective experience in terms of physical mechanisms in the brain. Put another way, is it even possible to explain subjective, phenomenal experiences in terms of objective, physical mechanisms?

Many optimistic cognitive scientists argue that they will eventually be able to develop rigorous and satisfying explanations for subjective experience once they learn more about the mind and brain. Although consciousness seems mysterious at the moment, that just reflects the current lack of knowledge. There are now explanations for the physical mechanisms underlying heat and light, but there was a time when those concepts also seemed mysterious and difficult to explain in purely physical terms. Perhaps consciousness will also eventually succumb to the relentless progress of science.

Other cognitive scientists are not convinced by these analogies. After all, heat and light are physical constructs that can be directly measured. Subjective experience seems different. Although scientists can certainly measure the neural correlates of conscious experience, those neural correlates are not the experience itself. Science will continue to make progress in understanding the physical correlates of human experience, but in doing so, scientists will be answering only Chalmers's easy problems. The hard problem may not even be amenable to scientific investigation.

Consciousness and Intelligence

Another question that cognitive science aims to answer is whether consciousness is a necessary component of human-level intelligence. That is, if scientists build a computer program that exhibits human-level intelligence, will it necessarily also exhibit consciousness? Or could scientists build artificially intelligent zombies that are as smart as human beings but that don't have any kind of mental experience or qualia?

First, what does *human-level intelligence* mean? How would an AI researcher know when they've succeeded in building a system that is truly intelligent? The most popular answer is to apply the Turing test, which was proposed by the British mathematician and computer scientist Alan Turing in 1950. Imagine that you're typing messages back and forth with two different individuals. One of them is a normal human being, and the other is an artificially intelligent computer program. Your job is to try to figure out which is which.

You can ask any question you like, including questions about subjective feelings, but you never get to see the entities on the other side of the conversations. You see only their typed responses. If you can't determine which is the human being and which is the computer program, then the program would have passed the Turing test, and you could say it successfully exhibited human-level intelligence.

Now, suppose that such a computer program has been built and has passed the Turing test, and everyone agrees that it exhibits human-level intelligence. Furthermore, it claims to feel all the same emotions that human beings feel and to have subjective experiences just like humans do.

Does being able to pass the Turing test necessarily imply that the program is conscious? The philosopher John Searle presented a very famous argument that illustrates the possibility. It's called the Chinese room argument.

Imagine that a man is sitting in a room that has two windows. Someone hands him Chinese characters through one window, and he has to pass other Chinese characters out the other window. He doesn't know Chinese, but he has a giant book of instructions that tells him what characters to pass out given the characters that come in.

For the sake of argument, assume that the characters come out of the Chinese room at the same speed as a native speaker of Chinese would produce them. It might look like the person inside the room understands Chinese even though he clearly doesn't.

Searle claimed that the same argument applies to a computer that has been programmed to process Chinese. Even if it looks like it really understands Chinese, and even if it passes the Turing test, it's still just mindlessly following a set of instructions—just like the man in the Chinese room.

To really understand Chinese, one has to actively entertain conscious thoughts, choose which thoughts to convey as output, convert those thoughts into words and sentences, and take in and comprehend the thoughts of others based on what they say. Searle argues that a computer program can simulate consciousness by following a set of rules, but it fundamentally lacks the subjective experience of consciousness that sets humans apart. Searle's argument has been very influential, but it is by no means universally accepted.

Another famous thought experiment suggests that building artificial systems with humanlike experience should be possible. Imagine that neuroscientists have managed to build an artificial neuron out of silicon that behaves exactly like a real neuron. In fact, you could replace one of your neurons with one of these artificial neurons, and your brain would function exactly the same way. Now, imagine that you keep replacing neurons until you've replaced them all with artificial neurons.

Are you still conscious? Most people have the intuition that you would be. After all, if consciousness arises from the brain, and if the brain is functioning in exactly the same way, why would you no longer be conscious? And of course, if you are still conscious, then that implies that it's possible, in theory, to build an artificial system that exhibits consciousness. Scientists just need to get the details right.

Reading

Blackmore, S., and E. T. Troscianko. *Consciousness: An Introduction*. 3rd ed. Abingdon, Oxon: Routledge, Taylor & Francis Group, 2018.

Chalmers, D. J. *The Character of Consciousness*. Oxford: Oxford University Press, 2010.

Dennett, D. C. *Consciousness Explained*. London: Penguin, 1993.

Gazzaniga, M. S. "The Split Brain Revisited." *Scientific American* 279, no. 1 (1998): 50–55.

Koch, C. *The Feeling of Life Itself: Why Consciousness Is Widespread but Can't Be Computed*. Cambridge, MA: MIT Press, 2019.

21

Putting It Together: Unified Theories of Cognition

You may be familiar with attempts to develop unified theories in physics—theories of everything that try to integrate and explain every aspect of the physical universe, from gravity to strong and weak nuclear forces. This lecture explores efforts to develop something similar with the human mind. Unified theories of cognition try to integrate theories of perception, attention, learning, memory, and other aspects of cognition in a single unified framework. This is obviously a very challenging endeavor, but some real progress has been made.

Early Cognitive Architectures

Professor Allen Newell, one of the founding fathers of artificial intelligence and cognitive science, motivated the need for unified theories of cognition in his famous 1973 paper titled, "You Can't Play 20 Questions with Nature and Win." He pointed out that a lot of research in cognitive science consists of experiments that essentially answer yes-no questions about the mind. This approach is powerful, efficient, and effective. But Newell argued that relying exclusively on this kind of 20-questions approach also has a downside, particularly when the goal is to understand the human mind. The questions scientists answer become more and more specific, and so they risk never putting all the pieces together.

The approach that Newell and others have taken is to try to implement a so-called cognitive architecture—that is, a computer system that simulates how many different parts of the mind work together. Researchers can then try to simulate different experimental results using that architecture and then repeatedly revise it to overcome its shortcomings.

Newell worked on one of the first such cognitive architectures in the 1980s and 1990s. It was named Soar, and it formed the basis for a very influential book called *Unified Theories of Cognition*. Unfortunately, Newell died shortly thereafter, and although the Soar architecture is still very much alive and well, it has mainly been used to try to solve problems in artificial intelligence rather than trying to explain how the human mind works.

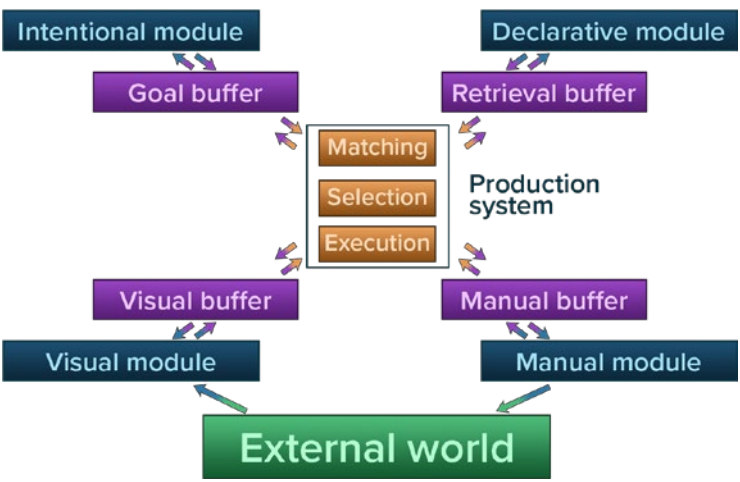
Overview of the ACT-R Theory

A different unified theory of cognition is called ACT-R. It isn't the only cognitive architecture out there, but it is among the most developed. Understanding how it works will give you a glimpse into what unified theories are like and will also provide insight into one particular unified theory of the human mind.

ACT-R was developed by John Anderson and his team at Carnegie Mellon University. It has been implemented in the form of a computational cognitive architecture, but the developers of ACT-R have explicitly tried to use it to simulate results from experiments on human beings. Their goal is to better understand how the human mind works.

You can download ACT-R and run it on your own computer. If you search for it in a browser, such as Google, one of the top results should be the ACT-R website at Carnegie Mellon University. From there, click on the Software tab, and you can download a stand-alone version of ACT-R for PC or Mac. The download also includes a folder with tutorials that you can work through.

This figure illustrates some of the main components of ACT-R and how they work together. ACT-R assumes that the human mind is organized into a set of separate processing modules, each of which performs a different function. The visual module receives visual information from the external world and identifies visual objects. The manual module controls the hands as they interact with the external world. The declarative memory module stores long-term memories and retrieves appropriate memories at appropriate times. Finally, the intentional module, which is also called the goal module, keeps track of the current goal so that the system can take steps toward achieving that goal.



Each processing module also has an associated buffer, which is used to communicate with the production system in the center of the diagram. So, most of the processing going on in each module is happening behind the scenes, separated from the rest of the system by a buffer. The only way the modules affect other parts of the system is by changing the contents of their buffer—that is, by using the buffer to present certain information to the rest of the system.

Each buffer can hold a single piece of information, which in ACT-R is referred to as a chunk. For example, one chunk might contain the information that 5 plus 2 equals 7. Another chunk might specify that your current goal is to solve an arithmetic problem. Another chunk might contain a description of what you're currently looking at through your eyes.

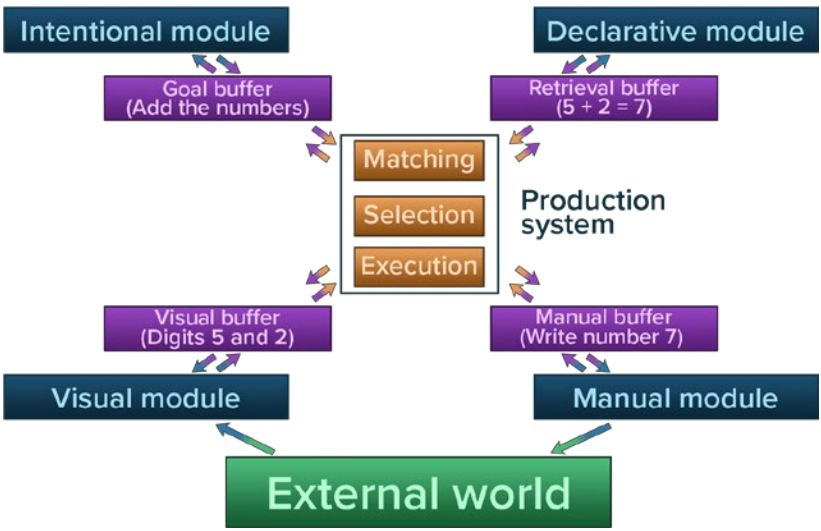
A module could perform many computations to determine what chunk should be in its buffer. For example, the visual module has to do many computations on the visual input before it can recognize an object, like a chair, or a person, like your friend. But you aren't aware of most of that processing. The only thing that the rest of the system knows is that you're looking at a chair or at your friend, because that's the chunk that ends up in the visual module's buffer. Likewise, the declarative memory module stores a vast knowledge base of information about the world but only retrieves a single chunk that best matches whatever the central production system requests.

Finally, at the center of the diagram is arguably the most important piece of the entire ACT-R architecture: the central production system. It's a collection of associations between conditions and actions. These condition-action relationships are reflexive and automatic, and they're implemented as if-then rules. If the conditions of a given production rule are satisfied, then the actions specified can be taken.

The conditions correspond to the contents of the buffers of the different processing modules, and the actions correspond to changes to the contents of those buffers or requests to the modules to do something and return the result in their buffer.

For example, imagine that the visual buffer says you're looking at the digits 5 and 2, and the goal buffer says the goal is to add the numbers you're looking at. Then, a production rule might match against the chunks in the

visual buffer and the goal buffer and then send a request to the declarative memory module to retrieve the sum of 5 and 2 from long-term memory. The declarative memory module would then retrieve the knowledge that 5 plus 2 equals 7 as a chunk in its buffer. Then, back in the central production system, another production rule might test that the answer has been retrieved in the declarative memory buffer and then send a command to the manual module to write the number 7 on a piece of paper.



Chunks in all the buffers are determined based on input from the external world or from previous internal processing. Production rules match against the chunks in these buffers and then make changes to the buffers or request the modules to do something. The buffers then get updated and the cycle repeats. It's also worth noticing that the ACT-R theory assumes that many different processes are happening simultaneously and in parallel.

ACT-R also assumes that only one production rule can be executed at a time and that only a single chunk of knowledge can occupy a buffer at a time. So, ACT-R assumes that there's a bottleneck in the system in at least two places.

At a general level, ACT-R provides a theory about how goals, memory, vision, and motor control interact via modules, buffers, and a central production system. But the ACT-R theory also proposes specific ideas about the inner workings of each separate processing module. For example, how does the declarative memory module decide which chunk of information to retrieve at any given time?

It does so based on what ACT-R calls activation levels. Each chunk of knowledge in declarative memory has an associated activation level, which is the sum of two numbers: the chunk's base-level activation and its associative activation.

Base-level activation reflects how useful that chunk has been in the past. The more frequently and recently a chunk has been retrieved, the higher its base-level activation. Associative activation measures how relevant a chunk is to the current goal.

For example, if your current goal is to solve an arithmetic problem, then chunks of knowledge related to arithmetic will have high associative activation. If your goal changes to brushing your teeth or making breakfast, then the associative activation of chunks related to those goals will suddenly increase.

This theory has successfully explained a number of experimental results about human memory retrieval. Just like the declarative memory module has to decide which chunk to retrieve from long-term memory, the central production system at the center of ACT-R has to decide which matching production rule to execute at any given time. And it does so by keeping track of each production rule's utility.

Like chunk activations, production rule utilities are numbers that reflect how likely a given production rule is to be useful at any given time. A production rule's utility is calculated using the following equation: Utility = $(P \times G) - C$.

P is an estimate of the probability of achieving the current goal if the production rule is executed. G is the value of achieving the current goal. And C is an estimate of the cost required to achieve the current goal.

So, if achieving the current goal is very valuable and this production rule is likely to help in achieving that goal, then the production rule's utility will tend to be high. Conversely, if the goal is very costly to achieve, then that will tend to reduce the production rule's utility. These underlying parameters themselves actually evolve as the system learns the cost and the effectiveness of different production rules over time.

Using the ACT-R Count Model

Chunks

The first ACT-R model in the tutorial is called the count model. The rest of this lecture breaks down how ACT-R models human cognition by using this model to count from two to four.

```
(define-model count
    (add-dm
        (one ISA number number one next two)
        (two ISA number number two next three)
        (three ISA number number three next four)
        (four ISA number number four next five)
        (five ISA number number five)
        (first-goal ISA count-from start two end four))
```

The first line in the text file is "define-model count," which tells ACT-R that you're defining a new ACT-R model and that its name is count. Next, there's an "add-dm" command. The *dm* stands for "declarative memory," so this command adds chunks into ACT-R's declarative memory. The next six lines specify six chunks that should be added, and they are named one, two, three, four, five, and first-goal.

Each chunk has three parts. The first part is the chunk's name or identifier. The second part is the word *ISA* followed by a chunk type. The first five chunks here are "number" chunks, while the last chunk is a "count-from" chunk. The chunk type tells you what kind of information the chunk contains. For example, the number chunks in this model tell you what number comes after the number one, what number comes after the number two, and so on. And the count-from chunk tells you where to start counting and where to stop.

The third part of each chunk contains the chunk's actual information. It takes the form of a set of slot-value pairs. For example, the "one" chunk has two slot-value pairs. The first slot is "number," and its value is "one." The second slot is "next," and its value is "two." Essentially, this chunk encodes the knowledge that the next number after the number one is the number two. The other number chunks encode similar pieces of knowledge.

Finally, the "first-goal" chunk is of type "count-from." This chunk also has two slot-value pairs. The first slot is "start," and its value is "two," while the second slot is "end," and its value is "four." This chunk encodes the information that the first goal is to count starting from two and ending at four.

Now, you need to tell the program what to do with this information. The next line in the text file that specifies the count model reads "goal-focus first-goal." The goal-focus command tells ACT-R to put the first-goal chunk into the goal module buffer.

```
(define-model count
    (add-dm
        (one ISA number number one next two)
        (two ISA number number two next three)
        (three ISA number number three next four)
        (four ISA number number four next five)
        (five ISA number number five)
        (first-goal ISA count-from start two end four))
(goal-focus first-goal)
```

Remember, each module uses its associated buffer to communicate with the rest of the system. Also recall that the goal module is what maintains the current goal or intention. So, putting the first-goal chunk into the goal module's buffer tells ACT-R that its goal is to count from two to four.

The Production Rule: Condition Side

The count model contains three very simple production rules, which control ACT-R's actions. In the first one, the first line contains the letter *p*, which tells ACT-R that you're defining a production rule. And the word *start* tells ACT-R its name. So, ACT-R now knows that you're defining a production rule that is named start.

Recall that production rules are automatic if-then rules that take specified actions when certain conditions are satisfied. The conditions correspond to specific chunks being in specific buffers, and the actions correspond to changing the buffer contents or asking a module to perform some action.

The "if" side, or condition side, of the production rule is what appears above the arrow, and the "then" side, or action side, is what appears below the arrow. So for this rule, the condition side consists of the four lines that start with "goal" and end with "nil."

START PRODUCTION RULE

```
(p start
 =goal>
   ISA      count-from
   start    =num1
   count    nil
 ==>
  =goal>
    ISA      count-from
    count    =num1
  +retrieval>
    ISA      number
    number   =num1
 )
```

This specifies what kind of chunk needs to be in the goal module's buffer for this production rule's condition to be satisfied. The top line, "=goal>," tells ACT-R to look specifically in the goal module's buffer. The next line tells ACT-R what type of chunk to look for there. In this case, the chunk in the goal buffer should be of type count-from, indicating that your goal is to count, starting from a certain number.

What number? That's where the next line, "start =num1," comes in. The equals sign in ACT-R is like a prefix. The term that follows it will be a variable, meaning it can be assigned different values.

So, here, *start =num1* means the chunk in the goal buffer must have a slot named start that has some value. And whatever value is in the start slot of that chunk will be assigned to the variable *=num1*. For example, if the chunk in the goal buffer has the value "two" in the start slot, then that's what will be assigned to the variable *=num1*.

Finally, there's the line "count nil." That means that the count slot will contain the number that is currently being counted: first two, then three, then four. And *nil* is a special symbol in ACT-R that means "empty" or "blank." So, "count nil" actually tests that the chunk in the goal buffer does not have a value for the count slot, meaning it hasn't started counting yet.

So, the condition side of this production rule tests that there is a chunk in the goal buffer that is of type count-from, that has a value in its start slot, and that does not have a value in its count slot. If, and only if, those conditions are satisfied, then this production rule will match. The condition side also assigns the value in the chunk's start slot to the variable *=num1*.

The Production Rule: Action Side

The action side of the production rule appears below the arrow. There are two actions here, one that changes the contents of a buffer and one that asks a module to perform an action. The first three lines after the arrow tell ACT-R to make a change to the chunk in the goal buffer—that's why the first line says "=goal>." The "ISA count-from" line says that you're changing a chunk of type count-from. You already know that because you tested it on the condition side, so it's not strictly necessary to include that here.

START PRODUCTION RULE

```
(p start
  =goal>
    ISA      count-from
    start    =num1
    count    nil
 ==>
  =goal>
    ISA      count-from
    count    =num1
  +retrieval>
    ISA      number
    number   =num1
  )
```

The third line says that the value of the count slot in the chunk should be changed to whatever value is currently assigned to the variable *=num1*. And recall that, in the condition side, you assigned the value of the start slot to the variable.

Essentially, this production rule is going to copy the value of the goal chunk's start slot to the goal chunk's count slot. For example, if the start slot for this chunk has the value two, then the count slot's value will be changed from being empty, or nil, to the value two.

In the production rule's second action, the "+retrieval" tells ACT-R to ask the declarative memory module to retrieve a chunk from long-term memory and to stick that chunk in the declarative memory module's buffer. Specifically, you want declarative memory to retrieve a chunk that is of type "number" with a value in its number slot that matches whatever is assigned to the variable *=num1*. And since *=num1* was assigned the value of the goal chunk's start slot, this is asking ACT-R to retrieve a number chunk corresponding to that value.

Note that the action side of the production rule never changes the value of a variable; it just uses whatever value was assigned on the condition side.

Other Production Rules: Increment and Stop

The are other two production rules. The increment production rule tests that you've retrieved a number chunk in the declarative memory module's retrieval buffer that matches the current count that you've reached, which is stored in the count slot on the goal chunk. It also tests that the current count is different from the number you're trying to count to, because otherwise you would be done counting.

INCREMENT PRODUCTION RULE

```
(p increment
   =goal>
      ISA      count-from
      count    =num1
     -end      =num1
   =retrieval>
      ISA      number
      number   =num1
      next     =num2
==>
   =goal>
      ISA      count-from
      count    =num2
   +retrieval>
      ISA      number
      number   =num2
   !output!    (=num1)
   )
```

If these conditions are satisfied, the rule updates the count value to be the number that the retrieval chunk tells you comes after the current count number. It also asks the declarative memory module to retrieve a new chunk that will tell you what number comes after that one. The last line tells ACT-R to print out the current count number so that the user can see the progress.

The stop production rule tests if the current count number is the same as the end number that you were trying to count to. If it is, then you're done counting, and the rule removes the goal and prints out the number.

STOP PRODUCTION RULE

```
(p stop
 =goal>
    ISA       count-from
    count     =num
    end       =num
 =retrieval>
    ISA          number
    number    =num
==>
 -goal>
 !output!   (=num)
 )
```

Running the Act-R Model

If you run this model, you can see all the specific steps that this ACT-R model takes as it counts from two to four. After setting the goal to be to count from two to four, the system fires the start production, which sets the counter to the start number specified in the first-goal chunk. It also requests retrieval of a chunk that specifies what number comes after that. It then retrieves the number two chunk from long-term memory and stores it in the declarative memory module's retrieval buffer.

Then, the increment production fires and prints out the number two, increments the count to three, and requests retrieval of a chunk that specifies the next number after three. Once the three chunk is retrieved, the increment rule fires again, this time printing out three, incrementing the count from three to four, and requesting retrieval of the next number chunk. This continues until the count reaches four, which is what the first-goal chunk specified as the end of the count. Then the stop production fires, the last number is printed, and the goal is removed.

0.000	GOAL	SET-BUFFER-CHUNK GOAL FIRST-GOAL NIL
0.000	PROCEDURAL	CONFLICT-RESOLUTION
0.000	PROCEDURAL	PRODUCTION-SELECTED START
0.000	PROCEDURAL	BUFFER-READ-ACTION GOAL
0.050	PROCEDURAL	PRODUCTION-FIRED START
0.050	PROCEDURAL	MOD-BUFFER-CHUNK GOAL
0.050	PROCEDURAL	MODULE-REQUEST RETRIEVAL
0.050	PROCEDURAL	CLEAR-BUFFER RETRIEVAL
0.050	DECLARATIVE	start-retrieval
0.050	PROCEDURAL	CONFLICT-RESOLUTION
0.100	DECLARATIVE	RETRIEVED-CHUNK TWO
0.100	DECLARATIVE	SET-BUFFER-CHUNK RETRIEVAL TWO
0.100	PROCEDURAL	CONFLICT-RESOLUTION
0.100	PROCEDURAL	PRODUCTION-SELECTED INCREMENT
0.100	PROCEDURAL	BUFFER-READ-ACTION GOAL
0.100	PROCEDURAL	BUFFER-READ-ACTION RETRIEVAL
0.150	PROCEDURAL	PRODUCTION-FIRED INCREMENT
TWO		
0.150	PROCEDURAL	MOD-BUFFER-CHUNK GOAL
0.150	PROCEDURAL	MODULE-REQUEST RETRIEVAL
0.150	PROCEDURAL	CLEAR-BUFFER RETRIEVAL
0.150	DECLARATIVE	start-retrieval
0.150	PROCEDURAL	CONFLICT-RESOLUTION
0.200	DECLARATIVE	RETRIEVED-CHUNK THREE
0.200	DECLARATIVE	SET-BUFFER-CHUNK RETRIEVAL THREE
0.200	PROCEDURAL	CONFLICT RESOLUTION
0.200	PROCEDURAL	PRODUCTION-SELECTED INCREMENT
0.200	PROCEDURAL	BUFFER-READ-ACTION GOAL
0.200	PROCEDURAL	BUFFER-READ-ACTION RETRIEVAL
0.250	PROCEDURAL	PRODUCTION-FIRED INCREMENT
THREE		
0.250	PROCEDURAL	MOD-BUFFER-CHUNK GOAL
0.250	PROCEDURAL	MODULE-REQUEST RETRIEVAL
0.250	PROCEDURAL	CLEAR-BUFFER RETRIEVAL
0.250	DECLARATIVE	start-retrieval
0.250	PROCEDURAL	CONFLICT-RESOLUTION
0.300	DECLARATIVE	RETRIEVED-CHUNK FOUR
0.300	DECLARATIVE	SET-BUFFER-CHUNK RETRIEVAL FOUR
0.300	PROCEDURAL	CONFLICT RESOLUTION
0.300	PROCEDURAL	PRODUCTION-SELECTED STOP
0.300	PROCEDURAL	BUFFER-READ-ACTION GOAL
0.300	PROCEDURAL	BUFFER-READ-ACTION RETRIEVAL
0.350	PROCEDURAL	PRODUCTION-FIRED STOP
FOUR		
0.350	PROCEDURAL	CLEAR-BUFFER GOAL
0.350	PROCEDURAL	CLEAR-BUFFER RETRIEVAL
0.350	PROCEDURAL	CONFLICT-RESOLUTION
0.350	----------	Stopped because no events left to process

Obviously, counting from two to four seems pretty simple on the surface, but implementing a real working system that performs the task requires spelling out a lot of details. In short, you're forced to be completely explicit, which is very helpful to cognitive scientists who are trying to understand the nuts and bolts of the human mind.

Another very important observation is that unified theories like ACT-R specify the architecture of cognition but not all the knowledge contained in declarative and procedural memory. The assumption is that different people have unique collections of knowledge but that the underlying architecture is relatively similar in normal, healthy adults.

Reading

Anderson, J. R. *How Can the Human Mind Occur in the Physical Universe?* New York: Oxford Academic, 2007. https://doi.org/10.1093/acprof:oso/9780195324259.001.0001.

———. *The Architecture of Cognition*. New York: Psychology Press, 2013.

Laird, J. E., C. Lebiere, and P. S. Rosenbloom. "A Standard Model of the Mind: Toward a Common Computational Framework across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics." *AI Magazine* 38, no. 4 (2017): 13–26. https://doi.org/10.1609/aimag.v38i4.2744.

Newell, A. *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press, 1994.

22

The Rapid Rise of Artificial Intelligence

Artificial intelligence has played a major role in the history of cognitive science, though the two fields have different goals. This lecture discusses those goals as well as some recent advances in artificial intelligence and machine learning that have had a dramatic influence on theories of the human mind.

Related Fields with Different Goals

Artificial intelligence (AI) and cognitive science have a unique relationship. The two fields are intimately related, and many people would consider AI to be a specific branch of the more general field of cognitive science. There's some truth to that, but most researchers in AI have somewhat different goals than researchers in the field of cognitive science.

One way of think about it is that AI is fundamentally a branch of engineering in which the goal is to build systems that are as intelligent as possible, regardless of the underlying mechanism. AI researchers can, and often do, take inspiration from what is known about the human mind and brain, but the ultimate goal is functionality. AI research aims to build a computational system that works well, even if the way it works is fundamentally different from the way human beings solve the same task.

In contrast, cognitive science is a branch of science in which the ultimate goal is explanation—that is, to understand and explain how nature works and, specifically, how the human mind and brain work. And so, cognitive scientists want to develop models that behave the way humans do, including exhibiting the same imperfections that humans exhibit.

But even though the goals of AI and cognitive science are somewhat different, developments in AI have had a profound influence on cognitive science. It's important to note that AI is evolving so rapidly that by the time you complete this course, it's very likely that the field will have changed dramatically.

Development of Deep Learning

Deep neural networks are networks of artificial neurons and synapses that have many layers. They are trained using massive numbers of input-output pairs. Given such input, a supervised learning algorithm like backpropagation gradually changes the strengths of all the synapses until the network can produce a reasonably accurate output pattern for each input pattern.

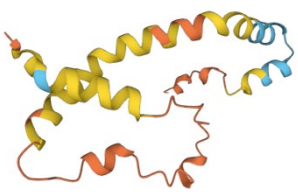
For instance, take the AI art generator made by the company OpenAI. It's known as DALL-E, after the Spanish painter Salvador Dali. Given any prompt, DALL-E can compose a work of art that resembles the content of the prompt with amazing accuracy. The results are very realistic and often all but indistinguishable from art made by a real person.

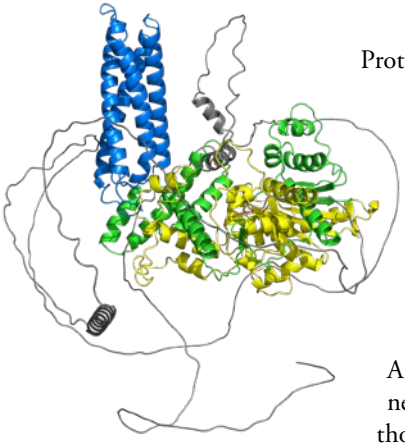
DALL-E is a generative model that uses a combination of deep learning techniques to generate art. The model works by training on a large dataset of images and then generating new images based on that training. Specifically, DALL-E uses a combination of a transformer-based language model and a generative adversarial network to create its images. You may recall transformer models from the discussion about modeling language in lecture 5. This is the technology that underlies ChatGPT and its various versions.

In the case of DALL-E, the transformer-based language model allows it to use language to guide the image-generation process. Users can provide textual descriptions to DALL-E, which then uses the language model to generate a set of features that represent the objects, scenes, and concepts described in the text.

DALL-E's generative adversarial network, or GAN, then generates the actual image from the set of features generated by the language model. The GAN is made up of two networks: a generator and a discriminator. The generator network creates images based on the feature set generated by the language model, while the discriminator network attempts to distinguish between the generated images and real images from the training dataset. The two networks are trained together in an adversarial manner, where, like a game, each network is pitted against the other. The generator attempts to create images that fool the discriminator, and the discriminator attempts to identify which images are real and which are generated. Over time, this process results in the generation of increasingly realistic and sophisticated images.

Another impressive application of deep learning is a system called AlphaFold. It was developed by the company DeepMind and can be used to predict the three-dimensional structure of proteins from their amino acid sequences.





Proteins are complex molecules made up of chains of amino acids that fold into specific three-dimensional shapes. The shape of a protein is critical to its function, and so understanding protein folding is essential for understanding many biological processes. It's also very important in the development of new drugs.

AlphaFold works by training neural networks with many layers on hundreds of thousands of input-output pairs. The inputs are amino acid sequences, and the outputs are the experimentally determined three-dimensional shapes that they fold into. After training, the system can be used to predict the three-dimensional structure of new amino acid sequences on which it wasn't trained. And it does so extremely well with very high accuracy.

AlphaFold has the potential to accelerate drug discovery by enabling researchers to design drugs that target specific proteins more effectively. It could also lead to a better understanding of the structure and function of proteins in the human body and their role in disease.

Deep Reinforcement Learning

Another important development is work that combines deep learning with reinforcement learning. A reinforcement learning agent learns by exploring the environment, taking actions, and receiving feedback on its actions in the form of rewards or punishments. The goal of the agent is to learn to select actions that maximize the expected long-term cumulative reward over time.

Reinforcement learning and deep learning have complementary strengths. Reinforcement learning is really good at learning what actions to take given the current state of the world. But it doesn't say anything about how the state of the world should be represented. Nor does it learn anything about state representations.

Deep learning, on the other hand, is great at learning to accurately represent the state of the world. A deep learning network can extract and represent detailed features of the world as it is. And those are exactly the kinds of features that a reinforcement learning system needs to know about. Deep learning gives reinforcement learning an accurate starting point to work from. Given this complementarity, a number of AI researchers have explored combining the two approaches in what is typically called deep reinforcement learning.

Deep reinforcement learning uses deep neural networks to learn powerful state representations and to try to predict the value of different actions from those states. Reinforcement learning then chooses the next action to take. Once it has chosen an action, it compares the value of the chosen action with the value that the deep neural network predicted before. By comparing the actual values of the actions against its previously predicted values, reinforcement learning is able to compute reward prediction errors, which are used to change the weights in the deep neural network so that, next time, it will predict the value of each action just a little better.

Eventually, after hundreds of thousands or even millions of trials, the deep neural network learns state representations that allow it to accurately predict the value of different actions, and reinforcement learning learns powerful control strategies that maximize long-term cumulative reward.

Deep reinforcement learning has been used to achieve breakthroughs in many areas, including game playing, robotics, and natural language processing. For example, consider AlphaGo, developed by Google's DeepMind.

AlphaGo uses deep reinforcement learning to play the ancient Chinese board game Go, which is considered to be one of the most complex board games in the world. Go has a much larger space of potential moves to search through than most other board games, including chess, making it more difficult for a computer to evaluate potential moves.

AlphaGo combined a 12-layer convolutional neural network with reinforcement learning and was trained to try to mimic moves made by human experts based on a database of about 30 million moves from historical games. Once it had gotten reasonably good, it was further trained by repeatedly playing a huge number of games against itself.

In October 2015, AlphaGo defeated the European Go champion in five straight games. The following year, it won four out of five games against Lee Sedol, widely considered to be one of the top five players in the world.

A new and improved version of AlphaGo called AlphaGo Master then began playing games online against some of the best professional players in the world. It went 60 and 0, including three victories over the world's top-ranked player.

Then, another new version of AlphaGo called AlphaGo Zero was trained without consulting any database of human moves. Instead, it just played millions of games against itself. After 3 days of training, it beat the original version of AlphaGo in 100 straight games. And after a few weeks of training, it was significantly better than AlphaGo Master. The developers hypothesized that mimicking human experts had actually handicapped the previous versions, and by excluding this training step, AlphaGo Zero could learn completely novel strategies that human beings had never come up with despite playing for thousands of years.

GPT Models

No discussion of AI would be complete without including the generative pretrained transformer (GPT) models. Most people have heard about ChatGPT by now. This publicly available AI system captured the world's imagination with its ability to perform a wide variety of tasks, including

- answering questions about almost any topic with an accuracy that rivals or exceeds human beings,
- composing reasonably well-written essays on virtually any subject,
- writing poetry and song lyrics,
- translating between almost any pair of human languages very accurately, and
- writing and debugging computer programs.

GPT models are a specific example of deep learning models. Like other deep neural networks, GPT models depend on many layers of artificial neurons and synapses to learn more and more abstract regularities from the inputs. Abstract regularities refer to general patterns that appear again and again and that correspond to concepts that the model learns over time.

In GPT models, the input words are represented in a high-dimensional semantic space. Think of this space as a big box of words, where no two words can occupy the same position. Words that have similar meanings are nearby in that space. For example, the words *wind* and *breeze* don't share any letters in common or sound anything alike, but they have similar meanings. And so, *wind* and *breeze* would be represented using similar numerical vectors in GPT models. These high-dimensional vectors that reflect the meaning of words are often referred to as word embeddings, because every word is embedded in this high-dimensional space near other words with similar meanings.

GPT models also incorporate positional information about the order of the words in the input. First, each position in the input is represented using a numerical vector that reflects the position of the word in the input: first, second, third, and so on. Then, that numerical vector for a given position is added to the embedding vector of the word at that position. The resulting vector is a combination of the word's meaning and its position, both of which are important in interpreting language.

Perhaps the single most important idea that led GPT models to be so much more successful than their predecessors was the so-called self-attention mechanism, which adds information about other words in the input. For example, consider this sentence: *The dog ran to the car when it heard it pulling up*. Humans immediately recognize that the pronoun *it* that is doing the hearing probably refers to the dog, while the other pronoun *it* that is pulling up probably refers to the car. But GPT models need to learn those kinds of dependencies. And that's what the self-attention mechanism does. After a lot of training, GPT models learn how much weight to attach to the other words in the input when processing a specific word.

How does it derive those weights in the first place? It's related to how GPT models are trained. In most cases, they are trained on billions and billions of excerpts from the internet, including Wikipedia pages, social media posts,

books, magazine articles, and anything else it can get its hands on. And its task is very simple: Given a sequence of a couple of thousand words, try to predict what word comes next.

With every new source of text, GPT models make millions of small changes to all their artificial synapses, which in turn leads the attention mechanism to do a better and better job of weighting the importance of other words in the input. Each layer in the deep network begins to extract more and more useful regularities from the input, and the system gets better at predicting what the next word is likely to be. And eventually, these GPT models get quite good at predicting the next word and knowing how each word relates to the others nearby—so good, in fact, that they can answer questions, write essays, translate between languages, and write computer programs.

Reading

Alammar, J. "The Illustrated Transformer." Blog post. June 27, 2018. https://jalammar.github.io/illustrated-transformer/.

Ford, M. *Architects of Intelligence: The Truth about AI from the People Building It*. Birmingham, UK: Packt Publishing, 2018.

Goodfellow, I., Y. Bengio, and A. Courville. *Deep Learning*. Cambridge, MA: MIT Press, 2016.

23

Cognitive Science in the Field

This lecture explores several areas where cognitive science has real-world applications, including the fields of education, design, law, and marketing. You'll learn why some study techniques are better than others for improving long-term retention of information. You'll see how modern designers incorporate many of the ideas and methods from cognitive science when creating intuitive interfaces for users. You'll also learn how developments such as the cognitive interview have improved eyewitness testimony in court cases. Finally, you'll see how artificial neural networks help tech companies to personalize marketing.

Impacts on Education

Cognitive scientists have learned a lot about how students learn and process information. One of the most important insights has to do with study techniques. If you want to learn and retain information for the long term, you might want to make your study sessions more challenging. Bob and Elizabeth Bjork at the University of California, Los Angeles, refer to such approaches to learning as desirable difficulties—desirable because they help you learn and retain information for long periods of time but difficult because they tend to make it harder to perform the task rather than easier.

For example, suppose that you're reading a book on a new topic that interests you. Most people would just read a chapter and hopefully retain some of the content. That kind of passive approach is relatively easy, but studies in cognitive science have demonstrated that it's much less effective than more active approaches, such as taking notes, summarizing the material in your own words, and writing down questions that you later use to test yourself. Implementing these kinds of active strategies is obviously more difficult than just passively reading, but that difficulty is desirable because it leads to better long-term retention.

Another example of a desirable difficulty is distributed practice, which essentially means spacing out your study. Doing this makes things a little more difficult because you tend to forget information between the study sessions, and so you have to get back up to speed. Nevertheless, numerous studies have demonstrated that the person who spaces out their study will have better long-term memory for the information than a person who crams all their studying into one long session.

There's also convincing evidence that testing yourself will help you remember the information better than restudying it will. In fact, testing is better than restudying even if you never see the correct answers from the test! One hypothesis for why that might be is that testing yourself forces you to retrieve information from your memory, while restudying doesn't. And of course, the ultimate goal is to be able to retrieve the information when you want it later. So, practicing retrieving the information is more helpful than just reencoding it.

These kinds of insights can also help teachers teach more effectively. For example, teachers who quiz their students regularly may find that it helps the students retain the material. And when the exams roll around, the students may feel better prepared.

Impacts on User-Centered Design

The goal of user-centered and user-experience design is to build products and interfaces that human beings find intuitive and easy to learn and use. This type of design depends on principles that cognitive science has discovered about how people perceive, learn, and process information.

To illustrate, consider the interface of modern smartphones. The ways that you interact with smartphones are natural and do not need to be learned. That's intentional. This kind of design significantly reduces the user's cognitive load. Common functions like calls and texts can be executed with your voice without having to type anything or open any apps.

Furthermore, opening an app typically requires only tapping it with your finger rather than having to remember and type a particular sequence of keys to invoke a program. That kind of interaction is fast and significantly reduces demands on memory.

Many aspects of interface design are also shaped by a very famous principle from cognitive science called Fitts's law, named after the researcher Paul Fitts. It states that the time it takes a person to move to a target depends on two factors: the distance to the target and the size of the target. More specifically, the time to complete a movement is proportional to the logarithm of the distance to the target and inversely proportional to the size of the target.

User-centered research often involves gathering data about the users' needs, goals, and behaviors and can include surveys, interviews, and usability testing.

In practical terms, this means that people can move more quickly and accurately to larger targets that are closer to them, while movements to smaller targets that are farther away take longer and are more error prone. Fitts's law is the reason the icons on your smartphone are relatively large and in nonoverlapping locations on your screen.

User-centered design typically involves running cognitive science experiments and collecting real data to determine which designs are the easiest for people to use. Designers often use the data they collect to develop personas, which are fictional representations of the different types of users of the interface or product.

Another key step in effective design is task analysis, which also comes from cognitive science. Cognitive scientists try to break down tasks into individual steps to figure out how people solve them. But in user-centered design, the goal of task analysis is to make sure that the interface or product is designed to support the users' tasks.

Task analysis has been used in health-care environments like hospitals and nursing homes to redesign medication management systems. Administering medication to hundreds of people a day is complicated and time-consuming, but getting it right is critical. A number of researchers have conducted task analysis to identify the steps involved in medication administration and to determine which steps are most prone to error.

Doing so has led to redesigned medication management systems that include things like color-coded medication carts, simplified medication administration records, and clear instructions for staff. These new systems are significantly more efficient, and they substantially reduce the number of medication errors.

Impacts on Legal Situations

Imagine that a person is injured in a car accident and that the case goes to court. In such cases, you often need to figure out whether a human being is to blame and, if so, who is at fault. But that's not always easy, and it typically

requires having detailed knowledge of so-called human factors, including how people perceive the world, how they learn and remember, how fast they can react, and the effects of distraction, stress, and fatigue.

David Krauss, who completed a PhD in cognitive psychology and neuroscience at UCLA, regularly consults with lawyers and serves as an expert witness in court cases. Using his detailed knowledge of cognitive science, he analyzes situations like car accidents to figure out how the human factors influenced the accident and who is most likely to blame.

In many legal situations, questions about the human mind and human behavior arise, including questions about the reliability of eyewitness memory. Recall from the discussion of episodic memory in lecture 8 that human long-term memory is malleable and prone to distortion. So, to improve the legal system and reduce the number of unjust verdicts, it seems like eyewitness reports need to be made more reliable.

One of the most important things that cognitive science can offer in this area is the cognitive interview, which is a technique used by law enforcement to enhance the accuracy of eyewitness memory. It was originally developed by Edward Geiselman at UCLA, Ronald Fisher at Florida International University, and a number of their colleagues. The key idea is to help the eyewitness retrieve and report as much information as possible about an event they have witnessed.

The cognitive interview incorporates four main strategies. The first is to reinstate the context. Human memories are strongly tied to the context in which they were acquired, and the features that were present at the time of encoding can serve as retrieval cues later on. For example, eyewitnesses might be asked to imagine the lighting, sounds, and weather conditions at the time of the event.

The second key idea is to encourage the eyewitness to report everything they can remember, no matter how trivial or seemingly unimportant it may seem. The goal is to retrieve information that might otherwise be overlooked or that might seem insignificant.

Third, the eyewitness is encouraged to recall the event in a variety of different orders, such as starting at the end and working backward or starting in the middle and moving to the end. This technique aims to help the eyewitness access more of their memory by changing the retrieval cues they use to remember the event.

Finally, the cognitive interview encourages the eyewitness to describe the event from different perspectives, such as from the viewpoint of another witness or from a different location. The goal here is to help the eyewitness access and report different details that they may have missed or forgotten about when they thought about the event only from their own original perspective.

Studies have found that the cognitive interview produces significantly more accurate memories than traditional police interviews do. And revisions to the cognitive interview have increased memory accuracy even more.

The cognitive interview exploits ideas from cognitive science that are known to enhance memory and reduce the influence of external factors that might distort or interfere with recollection.

Impacts on Personalized Marketing

Personalization in online marketing can improve the user experience by making ads more relevant; when a customer is recommended products based on their own browsing and purchasing history, it can make it easier for them to find what they're looking for. Furthermore, targeted and personalized marketing campaigns can be more effective at motivating users to engage and take action.

However, these methods also raise some concerns around privacy and data security. It is, therefore, very important for companies to be transparent about their data collection and use policies and to give users control over their data and how it is used.

But how do tech companies like Google figure out how to personalize ads? Specifically, what methods from cognitive science are being used and how?

One very important technique is called collaborative filtering, which involves building a model of a user's interests and the kind of products they like. Once you've built these user models, then you can compare the models of different users. If two users have similar interests and have liked or purchased similar items in the past, then they are likely to have similar preferences in the future as well.

Collaborative filtering has been used successfully in many recommendation systems, such as movie and music recommendations on platforms like Netflix and Spotify. It has also been used in e-commerce and social networking websites to suggest products and friends to users based on their past behavior and preferences.

Artificial neural networks like those discussed throughout this course are also regularly used in marketing and in the creation of personalized ads. For example, neural networks can be used to analyze customer data and identify patterns and similarities between customers. Doing so makes it possible to segment customers into different groups based on their characteristics, behavior, and preferences. These segments can then be used to tailor marketing messages and offers to each group separately in a way that is most likely to resonate with them.

Neural networks can also be used to predict customer behavior, such as purchasing patterns, product preferences, and response to marketing campaigns. Neural networks can even be used to create personalized content, such as product descriptions, email subject lines, and social media posts. By analyzing customer data and preferences, neural networks can generate content that appeals to each individual customer. And with the rise of generative pretrained transformer models like ChatGPT, machines can now even produce the first draft of the ads themselves.

Neural networks can also be used to optimize pricing strategies for personalized marketing. This often goes by the name of dynamic pricing. For example, airlines use dynamic pricing to adjust ticket prices based on demand and supply. Likewise, rideshare apps like Uber and Lyft regularly use data analytics and algorithms from cognitive science to monitor the demand for rides in real time. They consider factors such as time of day, location, weather, events, and holidays to predict the demand for rides in a given area. When demand for rides exceeds the available supply of drivers, the price of the ride increases. Dynamic pricing also incentivizes drivers to go to areas where demand is high so that they can earn more money.

Reading

Brown, P. C., H. L. Roediger III, and M. A. McDaniel. *Make It Stick: The Science of Successful Learning*. Cambridge, MA: Harvard University Press, 2014.

Loftus, E. F. *Eyewitness Testimony*. Cambridge, MA: Harvard University Press, 1996.

Norman, D. *The Design of Everyday Things*. Revised and expanded edition. New York: Basic Books, 2013.

Reason, J. *Human Error*. Cambridge: Cambridge University Press, 1990.

24

The Future of AI and Cognitive Science

As cognitive scientists develop a detailed understanding of how cognition works and how to simulate intelligence on machines, how will that affect the world? In particular, what ethical and safety concerns might the world face when scientists can build artificial intelligences that rival or even surpass human intelligence? And conversely, what benefits might advanced machine intelligence have for society? To close out the course, this final lecture considers these questions and what the future of cognitive science holds.

Transhumanism

In 2005, the futurist Ray Kurzweil wrote a book called *The Singularity Is Near: When Humans Transcend Biology* in which he outlines his hypothesis about what the future might look like as technology advances and scientists develop better and better models of the mind and brain. Specifically, he predicts that sometime around the year 2045, artificial intelligence will surpass human intelligence, leading to a "singularity"—a point in history after which human society will be fundamentally transformed.

Kurzweil argues that the power of computational devices is growing exponentially and will soon rival and eventually massively exceed the computational power of the human brain. Furthermore, he predicts that scientists will eventually be able to understand the brain in sufficient detail to reverse engineer it and ultimately build artificial brains that are dramatically more powerful than human brains are. And at that point, he speculates, people will begin augmenting their brains with nonbiological implants that increase their mental capacities. He argues that humans will ultimately transcend biology entirely, at which point they will be able to upload themselves into different computational substrates, alter their physical body in any way they like, and live more or less forever if they so choose.

Many of Kurzweil's ideas are reflected in the so-called transhumanism movement, which advocates for the use of advanced technology to enhance the human condition, including physical and cognitive abilities. Critics of Kurzweil's ideas often argue that progress isn't really exponential, and they point to counterexamples like space travel and oil production, where rapid developments eventually slowed down.

The goal of transhumanism is to transcend the limitations of the human body and mind and to achieve a post-biological state of existence with increased lifespan, enhanced intelligence, and improved physical abilities.

Other critics argue that, although the data that scientists are collecting about the human brain and body is growing quite rapidly, their understanding of how these systems actually work is growing much more slowly by comparison. But even critics typically agree that Kurzweil's analysis is thoughtful and well informed by current scientific advances.

Concerns in Education

One scary truth is that publicly available AI systems like ChatGPT and its successors can now do academic work with a quality that matches and often exceeds what human college students can do. In addition to writing excellent essays on virtually any topic, GPT systems can also translate between different languages, and their translations are rated extremely highly by native speakers. They can also write poetry and generate visual art, like paintings and computer graphics. They can even solve many problems from math and engineering.

Of course, these systems aren't perfect. They can and regularly do make mistakes. But they're going to get better, not worse, as time goes by. And there's currently no way for teachers to reliably distinguish academic work that was generated by a human being from work that was generated by an AI model. That raises a serious concern: Is there any way to ensure that students are actually doing the work they're being assigned? What if they're just plugging their assignments into a GPT model and then turning in the results, perhaps after looking over the output and making a few revisions?

Unfortunately, at present, there's no way for educators to prevent that kind of behavior. So what should educators do? What's the right way to think about this problem?

One idea is to tell students to do the work themselves and trust that they will. But now that these models are widely available, it's probably naïve to think that none of the students are using them. You could argue that the students who do use them are mainly harming themselves. After all, if they let these AI systems do their work for them, then they're obviously not going to learn

as much as their classmates, and they won't end up as well educated or as well prepared for the job market as those around them. And they're kind of wasting their tuition.

But it seems irresponsible for educators to wash their hands of the whole matter and just invite students to police themselves. Furthermore, in classes that are graded on a curve, comparing assignments that were produced by real students with assignments that were produced by an AI model is obviously very unfair to the students who did the work themselves.

So, if educators can't just expect students to always do the work themselves, what can they do? One approach could be to have students write essays in a system that tracks all revisions. For example, if a student writes a paper using Google Docs, the system will keep track of the entire history of how the document was written, including text insertions and deletions. Large language models like ChatGPT would write the entire essay from beginning to end with no revisions at all. So, the revision history of a student-written essay would look very different from the revision history of an AI-generated essay. Of course, students could find ways to get around this, but this approach might reduce the temptation to turn in AI-generated papers.

Perhaps the best solution for educators involves a combination of allowing students to use AI models in their work and designing assignments from which the students will still learn important skills and knowledge. For example, students could be asked to write an essay first and then use a GPT model to make suggestions for improvement. Alternatively, the students could be asked to revise and improve an essay that the AI model initially wrote.

Dangers of Perpetuating Social Prejudices

Another concern that has accompanied advances in AI and cognitive science is that AI systems may perpetuate and even amplify social prejudices, such as racial or gender bias. AI language models like ChatGPT and its successors are basically trained on language samples that are available on the internet. And because these sources often reflect existing biases and prejudices, there is a risk that the AI models will tend to exhibit exactly the same biases.

In one famous example, Amazon tried to train an AI model to identify highly qualified job applicants based on their resume. They trained the model by giving it thousands of previous resumes and telling it which were the resumes of people who had been hired and which were the resumes of people who had not been hired.

It turns out that the vast majority of technical jobs at Amazon had been filled by men in the past, and the AI model noticed this pattern. It therefore tended to favor the resumes of male candidates over the resumes of female candidates, even if both candidates were equally qualified for the job. Essentially, it noticed and then perpetuated an approach to hiring that was significantly biased. And of course, similar biases would show up if most of the previous hires were White rather than Black or lived in one part of town versus another or went to a particular school, and on and on. Fortunately, Amazon noticed the problem and scrapped the program. But it goes to show how AI models will tend to reflect the prejudices in society, because those prejudices tend to show up in the training data.

AI developers try to mitigate these kinds of problems in a number of ways. For example, they try to filter the training data to remove biased or offensive material before the model sees it. They also use reinforcement learning to reward the model for unbiased and true responses and to punish it for biased or untrue responses. Doing so requires human beings to decide whether the model's responses are true and unbiased or not, which is why this method is sometimes called reinforcement learning from human feedback, or RLHF. But doing so is both time-consuming and expensive. Furthermore, these approaches are not yet entirely effective, and many people have managed to elicit quite offensive output from AI models like ChatGPT despite the best efforts of the designers.

Economic Concerns

Advanced AI systems can perform many tasks that humans are regularly paid to perform, so these systems will undoubtedly eliminate jobs and displace employees. They are capable of processing vast amounts of data and generating complex outputs, which may make them useful for tasks like customer service, content creation, and even computer programming. They

also perform the tasks much faster than human beings do. And the quality of their work is often as good as the work of a human being and, at the very least, sufficiently good to satisfy the needs of a particular business. Finally, they don't need to be paid!

So, it seems inevitable in a capitalistic society that businesses are going to use these systems in many different ways to improve their bottom line. And of course, that means that jobs that used to be filled by human beings will no longer be available or, worse, that people whose jobs can now be done more cheaply by machines will be let go.

That said, it's important to keep in mind that AI systems are far from perfect. They regularly make mistakes and produce language output that is simply not true. That means that even businesses that use these systems will usually need to hire human beings to check the output for accuracy and revise it as necessary. So, while the demand for computer programmers and technical writers might go down, there may be new opportunities for people who are skilled in fact-checking and editing.

More generally, the development of advanced AI systems will also lead to the creation of other jobs. The same thing has happened repeatedly over the course of history. For example, the rise of machines and factory-based production during the Industrial Revolution led to the displacement of many traditional craftspeople and artisans. However, it also created new jobs in manufacturing, transportation, and other industries and helped spur economic growth and higher living standards.

How AI Can Benefit Society

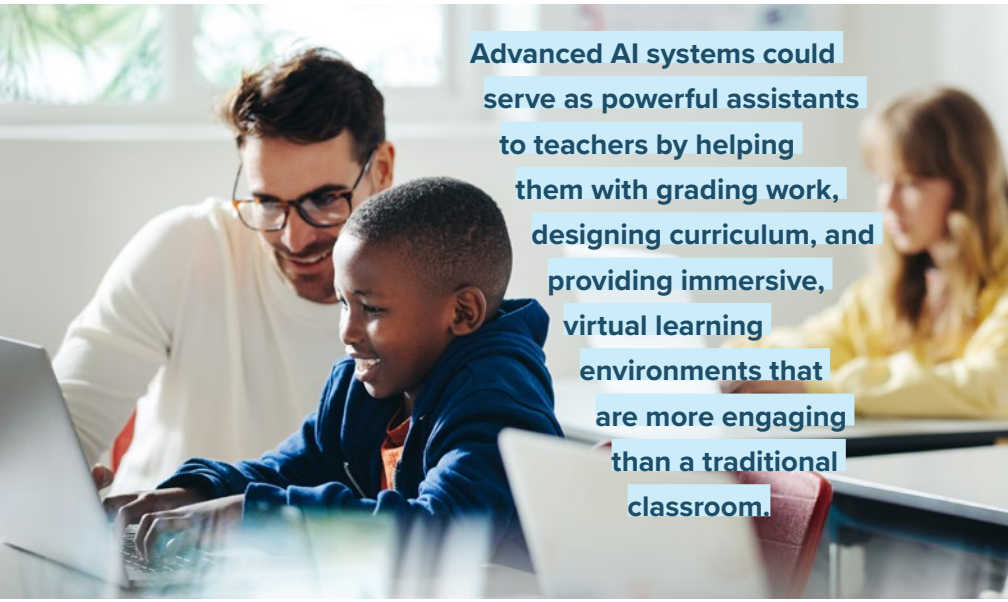
You may recall learning about deep convolutional neural networks in the discussion about computational models of vision in lecture 16. After being trained on millions of labeled images, these models can get very good at recognizing a variety of different objects in novel images.

The field of radiology and medical imaging is another domain that requires the ability to recognize particular features in images. Radiologists regularly have to read different types of medical images to decide if an abnormality is present that might need to be followed up. For example, they might need to

read a mammogram and try to find something that might be cancer. There's also an enormous database of previous mammograms for which the correct diagnosis was later confirmed by biopsy. And so, it's relatively straightforward to train a deep learning neural network to read mammograms and label them as needing a follow-up or not.

You would still want a human radiologist to be involved in the process, but a radiologist who is assisted by such an AI model is likely to be more accurate and efficient than one who is not. As Curtis Langlotz, a radiologist at Stanford put it, "AI won't replace radiologists, but radiologists who use AI will replace radiologists who don't!"

Advances in cognitive science could also have tremendous benefits in teaching and learning. For example, one problem that virtually all educators face is how to help the students who are struggling without boring the students who are excelling. What's needed is more personalized instruction that is tailored to the needs of the individual student.



A number of projects are exploring ways that machine learning could be used to analyze data on student performance and generate learning plans that are personalized to individual students. Furthermore, AI-powered tutoring systems could then provide students with instant feedback, guidance, and support, both in the classroom and at home.

Now that you've completed this course, hopefully you've gained some powerful insights into how your own mind works, from language and vision to memory and decision-making. And hopefully, you've also gained a new appreciation for the ways that insights from cognitive science are dramatically changing the world and will likely to change in the future.

Reading

Kurzweil, R. *The Singularity Is Near*. New York: The Viking Press, 2005.

Marcus, G., and E. Davis. *Rebooting AI: Building Artificial Intelligence We Can Trust*. New York: Vintage Books, a division of Penguin Random House, LLC, 2019.

Tegmark, M. *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York: Vintage Books, a division of Penguin Random House, LLC, 2018.

Villasenor, J. "How ChatGPT Can Improve Education, Not Threaten It." *Scientific American* (2023). https://www.scientificamerican.com/article/how-chatgpt-can-improve-education-not-threaten-it.