THE GREAT COURSES®

Topic
Philosophy & Intellectual History

Subtopic
Understanding the Mind

Philosophy of Mind: Brains, Consciousness, and Thinking Machines

Course Guidebook

Professor Patrick Grim
State University of New York at Stony Brook

PUBLISHED BY:

THE GREAT COURSES
Corporate Headquarters
4840 Westfields Boulevard, Suite 500
Chantilly, Virginia 20151-2299
Phone: 1-800-832-2412
Fax: 703-378-3819
www.thegreatcourses.com

Patrick Grim, B.Phil., Ph.D.

Distinguished Teaching Professor
of Philosophy
State University of New York
at Stony Brook

Patrick Grim is Distinguished Teaching Professor of Philosophy at State University of New York at Stony Brook. Graduating with highest honors in both Anthropology and Philosophy from the University of California at Santa Cruz, Professor Grim was named a Fulbright Fellow to the University of St. Andrews, Scotland, from which he received his B.Phil. He received his Ph.D. from Boston University with a dissertation on Ethical Relativism, spent a year as a Mellon Faculty Fellow at Washington University, and has been teaching at Stony Brook since 1976. In addition to being named SUNY Distinguished Teaching Professor, he has received the President's and Chancellor's Awards for excellence in teaching.

Professor Grim has published extensively on such topics as theoretical biology, linguistics, decision theory, artificial intelligence, and computer science. His work spans ethics, philosophical logic, game theory, philosophy of science, philosophy of law, philosophy of mind, philosophy of language, contemporary metaphysics, and philosophy of religion. Professor Grim is the author of *The Incomplete Universe: Totality, Knowledge, and Truth*; the co-author of *The Philosophical Computer: Exploratory Essays in Philosophical Computer Modeling*; the editor of *Philosophy of Science and the Occult*; and a founding co-editor of more than 20 volumes of *The Philosopher's Annual*, an anthology of the best articles published in philosophy each year. He has taught a course titled *Questions of Value* for The Teaching Company.

Professor Grim is perhaps best known for his critical logical arguments in the philosophy of religion and for his groundbreaking work in philosophical computer modeling. In this course, he draws from his broad interdisciplinary background to concentrate on philosophical issues of minds and machines,

brains and subjective experience, the phenomena of perception, and the mysteries of consciousness. ■

Acknowledgments

First and foremost, thanks to my wife, L. Theresa Watkins, for her work in shepherding lectures and supporting materials through multiple drafts. I have relied throughout on her ability to be scrupulous and critical while also creative and encouraging. I am also grateful to Christine Buffolino, Marcus Dracos, Rob Rosenberger, and Chris Williams for supplementary research and background support and to Marshall Weinberg and Marvin Levine for useful discussion. A special debt of gratitude for helpful feedback and discussion is owed to the students in my graduate and undergraduate courses in Philosophy of Mind. ■

Table of Contents

INTRODUCTION

LECTURE GUIDES

Table of Contents

Table of Contents

Philosophy of Mind: Brains, Consciousness, and Thinking Machines

Scope:

What is the relation between the brain and the mind? Is free will an illusion? Could a machine ever be creative? What is consciousness? The discipline known as *philosophy of mind* encompasses a range of questions about subjective experience, perception, intelligence, emotion, and the role of the mental in a physical universe.

Contemporary philosophy of mind is actively interdisciplinary. A broad range of disciplines is involved in the ongoing attempt to understand what minds are and how they work: psychology, neuroscience, cognitive science, artificial intelligence, computer science, and even robotics. This course highlights scientific results, provocative theories, and technological accomplishments in all of these fields in an exploration of what we know about our own mental functioning and what we do not. The overriding goal of the course is to develop a deeper philosophical understanding of our minds and of ourselves.

Each of the lectures focuses on a handful of intriguing questions in philosophy of mind: Is intelligence the same as IQ? Do minds function as parts or wholes? Do you see the same color I see? Do animals have a sense of self? Will our machines become smarter than we are? How can I know what other people think? Topics of investigation include color perception, body image, artificial intelligence, the structure of the adaptive brain, free will, self-identity, and current controversies regarding the nature of consciousness.

Thought experiments are an important conceptual tool. Here, science-fiction zombies, transporters, the inverted spectrum, Wittgenstein's beetles in the boxes, the Molyneux problem, Daniel Dennett's Chase and Sanborn, John Searle's Chinese room, and Ned Block's Chinese gym all play a role in philosophical exploration. The characteristics of injured brains offer a further conceptual resource, both real and tragic: the color-blind painter, the surprising phenomenon of blindsight, split answers from split-brain

patients, the isolation of autism, phantom limbs, prosopagnosia (the inability to recognize faces), the facts of psychopathology, and the strange case of Phineas Gage.

The history of ideas is woven throughout the course, from Greek concepts of the soul to the notion of mental substance in Descartes, sense-data in Locke and the Empiricists, Behaviorism in B. F. Skinner and Wittgenstein, and Functionalism in contemporary philosophy and cognitive science. Parallel ideas in philosophy, psychology, and the neurosciences are emphasized in examinations of eyewitness testimony, a mind in the world, the insanity defense, our understanding of other people, the "inner theater," and questions of conscious experience. The history of technology also plays a part, from ancient calculating devices to contemporary computers, from the golden age of automata to robots on Mars. Examples are frequently drawn from literature and the arts, with illustrations from Ovid's myth of Pygmalion through Shakespeare's *Macbeth* to science fiction films, such as *Blade Runner* and *The Matrix*.

Questions of mind are among the most hotly debated in philosophy today. This course outlines major positions in the debate in terms of their prominent contemporary defenders. Reductive Materialism and the new Dualism are considered using the work of Paul Churchland and Patricia Smith Churchland on one side and arguments from David Chalmers, Thomas Nagel, and Frank Jackson on the other. Functionalism and its critics are considered using the positions of Hilary Putnam and Daniel Dennett on one side with counterarguments from John Searle and Ned Block on the other. The philosophical debates spill over into other disciplines as well. The course features theory in computer science and the future of robotics by Rodney Brooks, Hans Moravec, Marvin Minsky, and Ray Kurzweil. It highlights work in psychology and neuroscience by J. J. Gibson, Francis Crick and Christof Koch, Oliver Sacks, Antonio Damasio, and V. J. Ramachandran.

The goal here, as in all philosophy, is conceptual clarification and rational argument. Philosophy has always thrived on controversy, and one goal of the course is to clarify core controversies in philosophy of mind by laying them out in terms of a range of intellectual options. What are the arguments behind Dualism and the considerations that have led most contemporary

thinkers to reject it? What is it that we see when we see color? Are there reasons to think that consciousness is *forever* beyond the reach of science? The aim of the lectures is always to open and articulate intellectual options, to capture the excitement of intellectual controversy, but never to lay down a single dogmatic position. Even where it is argued that one position is more plausible than another, the lectures attempt to present each side fairly enough so that the student may arrive at a different conclusion. One point about minds on which everyone agrees is that there is a great deal we do not yet know. It is important for continued progress to avoid closing off options too quickly.

The road map for the course starts with six lectures that lay out basic concepts, classical theories, and current hypotheses in philosophy of mind. Functionalism has emerged as a dominant trend in current research and continues as a theme in the next six lectures. This part of the course concentrates on perception, and our conceptions of ourselves and minds as they function in the world. Real robots play an interesting role in that exploration. In the third section of the course, we will focus on questions of intelligence—yours and mine—but also the idea of artificial intelligence. Strong conceptual connections through the course tie together an interdisciplinary examination of mind: Concepts introduced against a philosophical background may be illuminated in a later lecture with results from the neurosciences, compared with achievements in artificial intelligence, and then examined philosophically once again. The final six lectures of the course focus on subjective experience and the continuing mystery of consciousness.

The study of mind is inevitably double-edged. Here, as in all areas of inquiry, we use our minds to try to understand something. But in this area of inquiry, the object of the study—the thing we want to understand—is the mind itself. The fact that the mind is both subject and object of investigation opens wonderful opportunities for learning. The lectures sometimes rely on simple experiments, and because the subject of these experiments is mind, each listener will have all the equipment needed to participate. Auditory experiments are used at a number of points in the lectures to illustrate surprising aspects of the way in which we process sounds. Visual experiments appear in the lecture outlines, with further links to examples online.

Students can expect to gain from the course a rich understanding of rival theories and continuing philosophical controversies regarding minds and brains. Set in the context of intellectual history, that understanding will also be informed by some of the latest and most exciting work in psychology, the study of the brain, and information sciences. Everything we learn about minds makes them seem more interesting rather than less so. There is nothing more obvious than our own subjective experience, but there is also nothing more mysterious. ■

The Dream, the Brain, and the Machine

Lecture 1

The philosophy of mind represents intriguing questions regarding minds and brains that run from the mysteries of consciousness to prospects for thinking machines.

Many of our questions about brains, minds, and machines are not only questions for which we do not have answers, but questions for which we do not even have approaches for finding answers. What is consciousness? How could our three-pound brains possibly produce our rich experience of the world? Could a machine have subjective experience? These are the questions of philosophy of mind.

This introductory lecture outlines the kinds of topics to be covered using three examples: a particular dream, a specific brain, and the history of an intriguing machine. The dream is one that the young René Descartes had one night in 1619: a dream of the science of the future that would make clear the different realms of matter and mind. The brain is that of Albert Einstein, extremely similar to other brains—a little smaller than average, as a matter of fact—but with some unique characteristics as well. The machine is the Analytical Engine, designed by Charles Babbage in the mid-1800s. If the Analytical Engine had been built, it would have been a full-fledged computer constructed of steel and brass and driven by steam.

Photo by The Teaching Company.

Pythagoras considered himself a philosopher, not a mathematician.

This lecture also offers a road map of the course and illustrates the general approach to be used. The philosopher takes as input the range of different perspectives and the wealth of information available from all the arts and

sciences. That input, however, is merely a first step. A genuine understanding of the mind's role in the universe requires distinctly philosophical work as well: the disentangling of complex questions, the careful examination of alternative positions, and the search for rational argument and conceptual clarification.

This course covers both some amazing facts we know about the mind and some perplexing mysteries that remain. This first lecture offers a sample of topics to be covered in terms of three "exhibits":

- Exhibit A is a series of dreams that Descartes had in 1619.
- Exhibit B is Einstein's brain.
- Exhibit C is a steam-driven computer designed before the Civil War.

Exhibit A: In 1619, the young René Descartes envisaged a new science in a series of dreams. He dreamed of a philosophical unification of all the sciences, grounded in certainty and the solidity of mathematics. The core of that science was a radical distinction between minds and bodies. According to Descartes, the world is composed of two fundamentally different kinds of substances. Mass, energy, and the motion of atoms in the void belong to the physical substance of the universe. Feelings, ideas, and the taste of pineapple belong to the mental substance of the universe, as does Descartes' dream. Descartes' Dualism gives us the mind-body problem, an issue we will deal with in various forms throughout these lectures.

Exhibit B: What happened to Einstein's brain? Einstein's brain was "kidnapped" after his death, taken on a wandering journey from New Jersey to Kansas. Einstein's brain was both similar to, and different from, other people's brains. It was no bigger than the average brain. Its anatomy was different: An area associated with mathematical thought had taken over an area associated with language. Brain function and brain plasticity are themes that will reappear throughout this course.

Exhibit C: In the mid-1800s, Charles Babbage designed a computer. Babbage's Difference Engine was a special-purpose computer designed

to calculate logarithms. It was financed by the British government in the 1830s but abandoned before it was built. A Difference Engine was built in the 1990s from Babbage's plans and is now on display in the London Science Museum.

Babbage's Analytical Engine, designed well before Edison invented the light bulb, would have been a full-fledged, all-purpose computer made of steel and brass and driven by steam. Could a machine really be intelligent? Could it be creative? Could it be conscious? One way these lectures will look at minds is in terms of thinking machines.

History offers a tantalizing link between Descartes' dream and the concept of mechanical minds: the legend of Descartes' robot daughter. It is a story of minds, machines, dreams, and emotions. Philosophy began in a general "love of wisdom."

As time went on, philosophy progressively gave birth to mathematics, astronomy, physics, anthropology, sociology, linguistics, and psychology as specialized scientific disciplines. Once we think we have the techniques needed to answer specific questions, they become scientific questions rather than philosophical questions. Philosophy continues to tackle the core questions that remain—questions that are not only unanswered but elusive in terms of methods for finding answers.

As time went on, philosophy progressively gave birth to mathematics, astronomy, physics, anthropology, sociology, linguistics, and psychology as specialized scientific disciplines.

The goal of philosophy of mind is a unified understanding of mind and its place in the world. The focus is on conceptual clarification, the disentangling of complex questions, and the careful examination of alternative approaches. Philosophy's major tool is rational argument.

Throughout the course, we will lay out issues in terms of a range of intellectual options and alternatives. The first six lectures introduce basic concepts,

classical theories, and current hypotheses. The next six lectures follow the theme of Functionalism through issues of perception, self-conception, and minds in the world. A third section focuses on questions of intelligence, both natural and artificial. The final six lectures focus on subjective experience and the mystery of consciousness. A crucial prerequisite for the course is your mind, which will function both as the subject and the object of study. ■

Suggested Reading

Daniel Dennett, "Where Am I?" *Brainstorms*.

William Gibson and Bruce Sterling, *The Difference Engine*, 1st ed., pp. 1–71.

Douglas Hofstadter and Daniel Dennett, *The Mind's I: Fantasies and Reflections on Self and Soul*.

Questions to Consider

1. Do you think the mind is the same thing as the brain?

2. Do you think a computer could ever be creative?

The Mind-Body Problem

Lecture 2

In this lecture, I'll tackle the mind-body problem head on in terms of a position called *Dualism*. I want to start with five "obvious philosophical facts" or, at least, what seem to be obvious philosophical facts; these are things that just about everybody believes.

This lecture uses a logical principle from Leibniz and Descartes' assertion "I think, therefore I am" to present a major argument for Dualism. This lecture is based on five "obvious philosophical facts":

- You have a mind and a body;
- These normally work together;
- Your body is physical and, thus, publicly observable;
- Your mental life is essentially private; therefore,
- You have privileged access to the contents of your own mind.

The simplest position that makes sense of these philosophical facts is Dualism, developed in the work of René Descartes. This lecture uses a logical principle from Leibniz and Descartes' assertion "I think, therefore I am" to present a major argument for Dualism. The lecture then reveals a crucial conceptual problem that has convinced many philosophers that Dualism must be incorrect. In the Cartesian picture, physical things exist in physical space. The realm of the mental does not. But if mind and body are as radically distinct as Descartes says they are, how could they possibly interact?

The central problem will appear in a philosophical guise, a psychological guise, a neuroscientific guise, and in the guise of computer science. In this lecture, we will tackle the problem head-on, in terms of a position called *Dualism*.

We start with five "obvious philosophical facts"—things believed by just about everybody.

- Philosophical fact #1: You have a mind. You also have a body.
- Philosophical fact #2: Under normal circumstances, the mind and the body work together.
    - When you die, mind and body will not work together.
    - Some people think your mind will continue without your body, as a "disembodied" spirit or soul.
    - Some think that both your body and mind will simply cease to exist.
- Philosophical fact #3: Your body is physical.
    - It is composed of matter and occupies space.
    - As a result, the realm of bodily behavior is publicly observable.
- Philosophical fact #4: Your mental life is essentially private.
    - When you imagine a sunset, you alone can see it.
    - No one else can literally "feel your pain."
    - Unlike the physical realm, the mental realm is not publicly observable.
- Philosophical fact #5: You have *privileged access* to your own mental realm.

    - Suppose we ask Mary, "Does your knee itch?" Suppose we then ask John, "Does Mary's knee itch?" Mary says yes while John says no. Whose testimony should we trust?

    - Mary has privileged access to her own mental states.

    - Some have thought that privileged access extends to infallibility—that we cannot be wrong about our own mental states.

- Why do we think these obvious philosophical facts are true?

    - Some have proposed that they are the legacy of philosophical theories of the past.

    - Philosophers often do their best work by spelling out or making explicit concepts that are already part of common sense.

- Whether the obvious philosophical facts are, in fact, true is a question that will be open for further examination throughout these lectures.

If the obvious philosophical facts are true, what must the universe be like? The simplest theory of the universe that fits these facts is Dualism. According to Dualism, the universe is divided into two radically different halves. The physical realm contains all those things made of matter, which occupy space and are governed by the laws of physics. The mental realm contains those things that are essentially mental: hopes, emotions, imaginings, and consciousness. A can of pineapple is entirely physical. The taste of pineapple is something mental. This position is *Cartesian Dualism*, outlined in Descartes' *Meditations*, first translated into English as *Six Metaphysical Meditations: Wherein it is Proved that there is a God. And that Man's Mind is really Distinct from his Body.*

Descartes offered logical arguments that Dualism must be true. The best is that attributed to Descartes by Antoine Arnauld. The argument has two

related conclusions. The first conclusion is that your mind is in no way the same thing as your body or any part of your body. The second conclusion is that what is essential to you is not your body but your mind.

Crucial to the argument is a basic principle from Leibniz, the "indiscernibility of identicals":

- If two things are identical—if two things are the same thing—then anything true of one is true of the other.

- We should keep this principle in mind as we follow Descartes through the *Meditations*.

Recall Descartes' dream from Lecture One. Descartes sought complete certainty. His method was to try to doubt everything. If anything could be genuinely certain, it would be impervious to doubt. Let us follow Descartes in trying to doubt everything. Perhaps everything I've ever been told was a lie and nothing is as it seems. Perhaps even other people don't exist. In the *Matrix* movies, human beings are fed false subjective experiences, deceived into thinking that they are living normal lives. Suppose a race of robots, a master computer, a mad doctor, or an "evil demon" is trying to deceive me in any way it can.

The next crucial step in this process of doubting takes two forms. Is there anything I cannot doubt? Yes, that I am doubting. Were I to doubt that I was doubting, I would still be doubting. Is there anything I cannot be deceived about? Yes, that I am thinking. Were something to deceive me into merely thinking that I am thinking, I would still be thinking. Two conclusions seem to follow:

- If I am doubting, I must exist. If I am thinking, I must exist. *Cogito, ergo sum*: I think, therefore I exist.

- If I doubt, I must have a mind. If I think, I must have a mind. Therefore, I cannot doubt that I have a mind.

When the pieces of the argument are put together, they entail the conclusion that Dualism must be true. Leibniz's principle is: If two things are identical, everything true of one must be true of the other. I can doubt that I have a body or any part of a body. I can even doubt that I have a brain—maybe that is part of the illusion. I cannot doubt that I have a mind. There is, therefore, something true of my mind that is true of no part of my body: I cannot doubt that I have it. It follows by Leibniz's principle that my mind cannot be my body or any part of my body. My mind cannot be my brain.

Cartesian Dualism tells us that the mental and the physical are essentially different kinds of things. Physical things, Descartes says, are always extended and occupy space. Mental things do not have physical dimension in the same way. Dualism is the simplest theory that accords with our obvious philosophical facts, and we have a logical argument for it.

But Dualism also has a central philosophical problem. According to Cartesian Dualism, the mental and the physical are entirely different realms. One is a realm of things that obey physical laws and occupy space. Another is a realm of ideas, sensations, and feelings that don't even exist in space. If those realms were entirely distinct, it would seem that nothing mental could cause anything physical, and nothing physical could cause anything mental.

But we know that the mental does affect the physical: Our desires result in physical behavior. We know that the physical does affect the mental: Physical events in the world affect our beliefs and feelings. The "completely separate realms" view of Dualism must, therefore, be wrong. Gilbert Ryle referred to the Cartesian view of mind as the "myth of the ghost in the machine." Can we find a way out of this dilemma? Here is a suggestive analogy, the spirit of which will reappear in later lectures:

- Your fist is not identical to your hand. You can make a fist and then open your hand: Your fist has ceased to exist, but your hand has not. By Leibniz's law, they cannot be identical.

- Yet there is a sense in which your fist isn't something extra. To make a fist is just to put your hand in a particular position. ■

Suggested Reading

René Descartes, *Meditations on First Philosophy*, First and Second Meditations.

Gilbert Ryle, *The Concept of Mind*, chapter 1.

Questions to Consider

1. Our fifth "obvious philosophical fact" was that you have privileged access to the contents of your own mind. Some people have taken this to mean that you cannot be wrong about the contents of your own mind. What do you think? Can you be wrong about whether you are in pain? Can you be wrong about what you believe?

2. Consider this thought experiment: We haven't proven anything like this, but suppose the following beliefs turned out to be inconsistent. In order to remain consistent, you have to abandon at least one. Which of these beliefs would you keep, which would you reject, and why?

- The mental and the physical are radically different aspects of reality.
- The physical and the mental are causally linked.

Brains and Minds, Parts and Wholes

Lecture 3

However, merely stating that the mind has a physical understructure in the brain doesn't answer the major questions about the relation between mind and brain. It doesn't tell us what the physical understructure is. It doesn't tell us in precisely what ways aspects of our mental lives are linked to aspects of our brains.

The brain and the mind seem to work in parallel: The brain is the physical understructure of the mind. That fact suggests a strategy for investigation. We should be able to find out things about the brain by seeing how the mind works. We should be able to find out things about the mind by seeing how the brain works. This lecture offers a historical view of a number of things we have learned about minds and brains using precisely this strategy.

In 1848, a railroad foreman named Phineas Gage suffered a horrible accident: An iron rod was blasted through his head. Amazingly, Gage survived the accident, but he underwent a radical personality change. Together with the story of phrenology, the strange case of Phineas Gage is used to explore the history of a basic question about mental faculties and the brain: Does the brain function as a thing of distinct parts or as a unified whole?

If mind and brain are correlated, we will be able to learn things about the brain by studying the mind, and we will be able to learn things about the mind by studying the brain. This lecture also considers moral implications from some of the facts of interaction.

The accident suffered by Phineas Gage presents a classic case used to explore the question of mind-brain interaction. In September 1848, Gage was working on the railroad when a blasting charge sent an iron rod through his head. Gage showed no evidence of speech or memory problems, but he experienced a radical change in personality. Previously, he had been "the most efficient and capable" man in the railroad's employ. After the accident,

Gage was described as fitful, irreverent, profane, impatient, and obstinate yet capricious and vacillating.

In reconstructions using Gage's skull, Hanna Damasio and her students concluded that the rod passed through the ventromedial region of the frontal lobe. Such cases raise questions on both sides of the issue of mind-brain interaction. One set of questions relates to how brains function. Another set involves the character of our mental lives and moral experience.

Is the mind one thing or many? In one picture, the mind is a thing of distinct parts or mental "faculties" located in different parts of the brain. Another picture is an image of mind as a homogenous whole. William James noted the unity or "stream of consciousness." In that picture, we should perhaps expect to find a holistic brain of undifferentiated "mind stuff" that does everything everywhere.

Is the mind one thing or many? In one picture, the mind is a thing of distinct parts or mental "faculties" located in different parts of the brain. Another picture is an image of mind as a homogenous whole.

Which view is right? The case of Phineas Gage played an important role in the historical debate. Francis Gall's phrenology of the early 1800s was based on the concept of localizing mental faculties and personality characteristics in distinct places in the head. Phrenology was a pseudoscience, unscientific even by the standards of the time. Because Gage's memory and linguistic abilities were intact, his case was used at the time as an empirical argument for a *holistic* view of the mind. Today, Gage's case is used as one piece of evidence among many on the other side. Gage's personality changes fit a *localization theory*, reconfirmed daily in CAT scans, PET scans, and functional MRIs.

What mental functions did Gage lose? Gage no longer behaved in socially appropriate ways, and his decision-making abilities were seriously impaired. Antonio Damasio's patient "Elliot," who suffered brain damage in the same area, shows similar personality changes. Elliot does well on all standard intelligence testing, with no evident impact on memory or linguistic abilities.

Tests also indicate a normal level of moral development. With regard to practical decisions in his own life, however, he has everything necessary for responsible action except the ability to put it into play. Elliot's is a case of *volitional dysfunction*. He knows right from wrong but behaves as if he does not.

These concepts are central to the history of the insanity defense. By the criterion adopted into American law in 1882, a defendant is "not guilty by virtue of insanity" if he or she did not know the nature and quality of the act or did not know that the act was wrong as a result of laboring under a defect of reason from disease of the mind.

A more modern criterion has both a "knowledge" and a "volitional" prong. Defendants are "not guilty by virtue of insanity" if they lacked substantial capacity to appreciate the criminality of their acts or lacked substantial capacity to conform their conduct to the requirements of the law as a result of mental disease or defect.

Phineas Gage and Damasio's Elliot might have a defense of insanity under the second criterion but not the first. Because of the volitional prong, the jury found John W. Hinckley, Jr., "not guilty by virtue of insanity" for the attempted assassination of Ronald Reagan in 1981. As a reaction to that decision, the second criterion has largely been abandoned; insanity is once again treated as a matter of what one knows.

Cases like those of Elliot and Phineas Gage also raise moral questions. We will examine the issue of free will in a later lecture. Many philosophers agree that we do act freely. But do people with brain damage like that of Elliot and Phineas Gage have free will? Antonio Damasio says no. Imaging studies show a high incidence of similar frontal lobe abnormalities in our prison population, with clear links to violent crime. Do these facts about minds and brains have implications for our system of criminal justice? ■

Suggested Reading

Rita Carter, *Mapping the Mind*, chapter 1.

Antonio Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain*, chapters 1–3.

Malcolm Macmillan, *The Phineas Gage Information Page*, http://www.deakin.edu.au/hmnbs/psychology/gagepage/.

Questions to Consider

1. Mind and brain seem to work together, yet we seem to have a single, unified stream of consciousness, despite the fact that different aspects of perception, judgment, and emotion are processed in different parts of the brain. How can that be?

2. Consider this thought experiment: You are on a jury in a murder case. The defense argues that the defendant suffered brain trauma from an automobile accident a year earlier; as a result, he no longer acts from his own free will.

3. Describe a piece of evidence that would convince you that the defense is right and the defendant does not have free will.

4. Describe a piece of evidence that would convince you that the defense is wrong and the defendant does act of his own free will.

The Inner Theater

Lecture 4

In this lecture, I'll turn to an everyday notion of how our minds work that's suggested by those philosophical facts. My topic isn't really a formal philosophical theory, but it is a common concept of mental processing—the concept of the "inner theater of consciousness."

What happens when you see? A naïve characterization is as follows: Something in the world reflects or refracts light, which enters your eyes and impinges on your retinas. From there, an image is sent to your brain and is projected onto something like an inner screen. Images in this "inner theater of consciousness" also seem to explain the phenomena of memory, imagination, illusion, hallucination, and dreaming. This characterization fits well with Dualism and accords with many of the "obvious philosophical facts" outlined in Lecture Two.

This lecture challenges the "inner-theater view" directly. A range of experiments in touch, hearing, and conscious willing seem to show that something much more complicated is happening in perception and volition. Logical arguments against the inner theater are even stronger, demonstrating that this naïve conception of our mental processing could not possibly explain what it is supposed to.

The purpose of this lecture is to outline and challenge an "inner-theater picture" of consciousness. We use the terms *conscious* and *consciousness* in a number of ways. In a limited sense, someone is *conscious* if he or she is not asleep or in a coma. We speak of someone as *conscious of* something when he or she responds to it. The sense that will be at issue in this lecture is *phenomenal consciousness*: the realm of subjective experience.

According to the inner-theater theory of vision, light enters your eyes and impinges on your retina, and an image is sent to your brain. What happens when you see is that the image from the external world is projected on something like an inner screen. The same pictures-in-my-head theory could explain a range of other mental phenomena, including memory, imagination,

and dreaming. The inner theater is also conceived as the place in which we make decisions and initiate movement. The picture of an inner theater is so culturally ingrained that we have come to think of it as "common sense." But there are reasons to think that major parts of the theory are simply not true.

Empirical evidence counters the inner-theater theory of perception. In the inner-theater theory, images should show up on the inner screen in the order they come in from exterior sources. The *cutaneous rabbit* is evidence of a much more complicated phenomenon. If you are given a series of 12 taps spaced 50 milliseconds apart at three distinct points—your wrist, halfway up the forearm, and your elbow—you will perceive them as a continuous series of evenly spaced taps. It will feel as if a small rabbit is hopping all the way up your arm, including hops at places that were not tapped at all.

How does your brain know to perceive taps in the middle before they have occurred at the end?

The *auditory rabbit* is a similar effect with sound. Although equally balanced clicks are sounded in two widely spaced speakers, you hear an even series of clicks moving across the space in between. These experiments create a problem for the inner-theater theory because the taps perceived in the middle appear in consciousness before the taps at the end. But the effect appears only when there are taps at the end. How does your brain know to perceive taps in the middle before they have occurred at the end? The only explanation seems to be that the illusion of taps in the middle is created after the entire series of real taps has taken place. That account, however, violates the A-B-C aspect of the inner-theater theory.

Related phenomena occur in visual perception. Movies and animations depend on the *phi phenomenon*, in which a series of still pictures is perceived as showing genuine motion. When dots of one color are flashed one after another at distinct points on a screen, the appearance is of a single dot that moves. If the color of the dots changes, the dot is seen to change color as it moves. How does your brain know to perceive dots in the middle moving in the right direction? How does it know to perceive a dot changing in color

toward the correct shade? A story of continuity must be constructed after the data of the endpoint are already in.

Empirical evidence also counters the "control-center" aspect of the inner-theater theory. In the 1980s, Benjamin Libet and his team timed both subjective and objective events in simple voluntary movement. On the subjective side, the researchers timed the points at which people reported they (a) were conscious of willing their arms to move and (b) were conscious of their arms moving. On the objective side, the researchers timed (a) a readiness potential in the brain that results in an arm movement and (b) the point at which the muscles in the arm actually contract. The results showed that the consciousness of willing came almost half a second after the readiness potential. The brain is already going full speed to produce a movement before you are aware of willing that movement.

Conceptual arguments against the inner theater are even stronger. Consider this thought experiment: Imagine a pirate standing with his arms crossed beside an open treasure chest. Can you honestly say whether he is missing any fingers, or how many buttons are on his coat, or what color the jewel is in the upper left-hand corner of the treasure chest? Perhaps imagining is not like looking at a picture in an inner theater.

We seem to think we know where the inner theater is—right behind our eyes. But historical evidence indicates that that notion is merely a matter of cultural indoctrination. The Egyptians carefully preserved many of the body parts for the afterlife but threw the brain away. Aristotle thought that the heart was the seat of reason. The function of the brain was merely to cool the blood. No aspect of your subjective experience tells you the location of the inner theater. It could be anywhere or nowhere in space at all.

If correct, the inner theater should work equally well for all senses. Visual images come in and are projected on an inner screen. Tastes come in and are projected as … what? Smells come in and are projected as … what? When you think about it, the theory really works only for sight.

The theory doesn't explain what is happening in perception and systematically cannot explain what is happening. In order for images on the screen to

amount to seeing, something would have to be watching the inner screen—the *homunculus*, or "little man." How does this inner being feel things? Taste things? See things? If we need an inner being to explain seeing, we will need another inner being to explain the first's ability to see, and another being inside the second, and so on. In the end, the theory appears ridiculous, impossible, and pointless.

The inner-theater demolition crew has done its work, but questions remain. Given that both experimental and conceptual evidence indicates that the theory is wrong, why is the inner-theater view so tempting? We import analogies from the outside. We know that we see something when we look at a picture; perhaps, then, seeing everything is like looking at pictures in our heads. We know that we can be fooled by pictures; maybe hallucination is being fooled by pictures in our heads. The analogy breaks down, however, under empirical and conceptual arguments. If this inner-theater theory is not how experience works, how does it work? We do not yet have a better theory to replace the inner theater. ■

Suggested Reading

Daniel Dennett, "The Nature of Images and the Introspective Trap," *Content and Consciousness.*

Daniel C. Dennett and Marcel Kinsbourne, "Time and the Observer: The Where and When of Consciousness in the Brain," *Behavioral and Brain Sciences* 15.

David I. Shore, *Measuring Auditory Saltation*, http://www.mohsho.com/dshore/sound_top.html (a version of the auditory rabbit).

University of North Carolina at Charlotte, http://www.philosophy.uncc.edu/faculty/phi/Phi_Color2.html (an interactive exploration of the color phenomenon).

Questions to Consider

This lecture explored attempts to demolish the concept of the inner theater. But the truth may be that parts of the concept work and others do not. In that light:

1. In what ways is imagining a pirate like seeing a picture of a pirate and in what ways is it different?

2. In what ways is memory like retrieving a picture from the files and in what ways is it different?

Living in the Material World

Lecture 5

In this lecture, I want to talk about some philosophical alternatives. If Dualism *isn't* true, what might be true instead? But let me first talk about some alternatives within Dualism.

Despite its initial plausibility, Dualism has been shown to have basic conceptual problems. A crucial difficulty is the *interaction problem*. If the mental and the physical realms are as radically distinct as Dualism says they are, how could they possibly interact in the ways we know they do? Historically, philosophers made some valiant efforts in the attempt to save Dualism, but to the contemporary eye, those attempts look futile. What are our intellectual options if we give up Dualism? Do we have more plausible options?

This lecture presents a range of intellectual options, from the idea that the universe is purely mental (Idealism) to the view that the universe is purely physical (Materialism). According to Reductive Materialism, mental phenomena—emotions, beliefs, feelings, and sensations—ultimately reduce to physical phenomena. By contrast, Paul Churchland and Patricia Smith Churchland's Eliminative Materialism proposes that mental phenomena don't exist. Concepts of belief and sensations are like concepts of witches and will be left behind as science progresses. Materialism has seemed particularly attractive to many scientists and philosophers; we will consider it both in terms of its promise and the conceptual problems it raises.

Library of Congress, Prints and Photographs Division, LC-USZ62-475.

Just as science now rejects the concept of witches, it will one day abandon concepts such as fear, belief, and hope.

Historically, a number of attempts were made to answer the interaction problem within Dualism. Descartes thought mind meets body in the pineal gland. Most parts of the brain form symmetrical pairs, but the pineal gland does not. Descartes thought mistakenly that animals do not have a pineal gland. The central question, however, is not where the mental and the physical interact, but how they possibly could.

The idea that Dualism is true clashes directly with the idea that the mental and the physical interact. To address this conflict, we could give up the idea that Dualism is true. Alternatively, we could give up the idea that a connection exists between the mental and the physical. At the time of Descartes and for a significant period afterward, the second route was seriously explored.

Epiphenomenalism holds that physical events do not cause mental events. Mental events nonetheless "float above" the physical events. Thomas Huxley argued for Epiphenomenalism.

Occasionalism holds that physical events do not cause mental events or vice versa. On every occasion, God makes both happen, producing an illusion of interaction. Nicolas de Malebranche was the classic proponent of Occasionalism.

The *interaction problem* boils down to how two distinct substances could possibly interact.

Parallelism holds that God sets up the universe from the beginning with a "pre-established harmony" between the physical and the mental. From that point, the two operate side by side, like two clocks wound up in parallel. Leibniz is the classic proponent of Parallelism. These attempts to shore up Dualism look increasingly desperate.

If Dualism is not true, what is? A number of philosophical viewpoints are major contenders for answering this question. The *interaction problem* boils down to how two distinct substances could possibly interact. One answer might be that perhaps there is only one kind of "stuff." This position is called *Monism*, and it comes in several different forms.

One version of Monism, *Idealism*, asserts that there is just one kind of stuff, and it is the mental stuff. What evidence do we have of the existence of anything physical? Only our sense impressions—and those are mental. Idealism leads directly to *Solipsism*: the view that there is no universe (and no people) beyond my mental realm. Idealism and Solipsism look like dead ends.

Another form of Monism is *Materialism*. There is only one kind of stuff, and it is the physical. But if everything is physical, then everything in the mental realm must ultimately be physical: sensations, pains, pleasures, belief, and love. The problem for the Materialist is to fill out that "ultimately."

Most Materialists have characterized themselves as *Reductive Materialists*, asserting that all mental phenomena ultimately reduce to physical phenomena. Reductive Materialists often claim that mental states reduce to brain states: chemical or neurological patterns in the brain. How plausible is Reductive Materialism? A central form of the position, known as *type identity theory*, argues that all types of mental states are identical to types of brain states.

If that argument is true, then two people who hold the same belief would be in the same brain state. But brains differ from person to person. We have no reason to think that the way your brain encodes a particular belief is identical to the way mine does. The theory also faces a more general problem: If pain is a brain state, what brain state is it? The closest people have gotten to answering this question is to say that to be in pain is to have C fibers firing. But because C fibers are carbon-based, it would follow that no silicon-based creature on some other planet could possibly feel pain. That result seems both intuitively wrong and ethically dangerous. It appears that pain, fear, love, and hope could be multiply instantiated. If so, Reductive Materialism's identification of these mental states with any particular physical state must be wrong.

Eliminative Materialism is another form of Materialism, championed by the philosophers Paul Churchland and Patricia Smith Churchland. According to Eliminative Materialism, mental phenomena do not really exist; there aren't really any fears, hopes, or beliefs. We will eventually develop a science of human beings that will cover everything about them. When we do, we'll find

that concepts of fears, hopes, and beliefs will not be part of this science. If Dualism is not a satisfactory theory, what might be? Are we out of options? ■

Suggested Reading

Patricia Smith Churchland, "Reduction and Antireductionism in Functionalist Theories of Mind," *Neurophilosophy*.

Paul Churchland, "Eliminative Materialism and the Propositional Attitudes," *Journal of Philosophy* 78.

U. T. Place, "Is Consciousness a Brain Process?" *British Journal of Psychology* 47.

Questions to Consider

1. Below is a list of positions mentioned in the lecture. On a scale of 1 to 10, from "not plausible" (1) to "very plausible" (10), rate each in terms of how plausible you think it is right now. You can change your mind later, of course.

Cartesian Dualism: The universe consists of two radically different substances.

1 2 3 4 5 6 7 8 9 10

Epiphenomenalism: The mind "floats" above the brain but has no causal effect on it.

1 2 3 4 5 6 7 8 9 10

Occasionalism: There is no causal link between the mental and the physical, but God intervenes on every occasion to make it appear that there is.

1 2 3 4 5 6 7 8 9 10

Parallelism: There is no causal link between the mental and the physical. It appears that a link exists because God set the two realms working in parallel at the beginning of the universe.

1 2 3 4 5 6 7 8 9 10

Idealism: Everything that exists is ultimately mental.

1 2 3 4 5 6 7 8 9 10

Solipsism: Everything and everyone that exists, exists only in my mind.

1 2 3 4 5 6 7 8 9 10

Materialism: Everything that exists is ultimately physical.

1 2 3 4 5 6 7 8 9 10

Reductive Materialism: The mental reduces to the physical. Mental states are a kind of brain state.

1 2 3 4 5 6 7 8 9 10

Eliminative Materialism: Our concepts for mental states—fears, hopes, beliefs—will be eliminated in an eventual science of human functioning, rejected as concepts of witches and humors are now.

1 2 3 4 5 6 7 8 9 10

A Functional Approach to the Mind

Lecture 6

In the last lecture, I promised a new and different approach to some of the central questions we've been tracking, a different approach to questions of the nature of mind, a different approach to the relation between the mental and the physical.

First, we'll explore Analytical Behaviorism using Wittgenstein's private language argument, together with his parable of the beetles in the boxes. This philosophical position parallels but is distinct from the psychological Behaviorism of B. F. Skinner. The Analytical Behaviorist claims that mental states—decision, love, or belief, for example—are not part of some inner mental realm but merely complex patterns of behavior.

Hilary Putnam's *multiple instantiation* argument is used to introduce a more contemporary successor to Behaviorism. Functionalism, which is currently the dominant position in philosophy of mind and the cognitive sciences, is the view that mental states are functional states of an organism: complex patterns that link inputs both to behavioral outputs and to other constellations of mental states.

How are we to understand the relation between body and mind, between the mental and the physical? It may seem that we have exhausted all conceptual possibilities, but two further positions take a radically different approach. Behaviorism was prominent at the mid-20th century, in both a philosophical and a psychological guise. Functionalism is currently the dominant position in philosophy of mind. The purpose of this lecture is to explain why.

Behaviorism and Functionalism offer another route out of both Dualism and Monism. From the beginning, the debate has been framed in terms of a question about the basic "stuff" of the universe. Why should we think that the important differences are a matter of stuff?

Wittgenstein's *private language argument* was crucial in changing the character of the entire debate. The *private language argument* is as follows:

- If Dualism were right, then my believing, seeing, imagining, and loving would be essentially inner and private, inaccessible to anyone else.

- But that very claim was expressed using words we all know: *believing*, *seeing*, *imagining*, *loving*.

- Words are learned by correcting incorrect uses and praising correct uses. We must have learned these words in that manner.

- But if Dualism were right, these things would be inner and private, inaccessible to anyone else. If so, we could never have learned these words.

- We did learn those words; therefore, Dualism must be wrong.

Wittgenstein also expresses the argument in terms of the parable of the beetles in the boxes. Suppose we each had a beetle in a box into which no one else could look. One day, I say, "My beetle is so … 'glorp.'" You say, "Mine is, too. No, it's more 'nikiniki' than 'glorp.'" Those words, Wittgenstein says, could never acquire meaning. If Dualism were right, the same would be true of all mental terms. Because mental terms do have meaning, Dualism must be wrong. If our language of mental states is not about some private inner experience, what is it about? One proposal is that talk of mental states is really a way of talking about behavior.

The philosophical form of this position is *Analytical Behaviorism*: the view that mental concepts are definable in behavioral terms.

In psychology, *Behaviorism* is used to refer to the theory of B. F. Skinner, famous for stimulus-response theory and particularly for schedules of reinforcement. But Skinner was far from consistent in talking about Behaviorism. Sometimes, he talked as if mental states did not exist. Only behavior exists. Sometimes, he took a methodological stance: The subjective "inside" is scientifically inaccessible. Sometimes, he seems to be an Analytical Behaviorist: Talk about mental states is a roundabout way of

talking about behavior. When you think about what people do when they are in love, you end up focusing on behavior. Perhaps that is all that love is.

Analytical Behaviorism faces a number of conceptual problems. If mental terms have publicly observable criteria, we should be able to list them, but no Analytical Behaviorist has ever been able to give an exhaustive list for any mental term. When people have attempted an analysis, they have ended up using other mental terms. No matter what behavioral equivalent is offered, it appears that someone will be able to add further context that makes it clear that the mental concept does not apply.

When you think about what people do when they are in love, you end up focusing on behavior. Perhaps that is all that love is.

Although philosophers have abandoned the letter of Behaviorism, its spirit lives on in Functionalism. Here again, philosophy and psychology have worked in parallel. Functionalism is at the core of contemporary cognitive science.

One of the prime motivations for Functionalism is the *multiple instantiation* argument. It appears that pain could have any of multiple instantiations: different physical structures in different kinds of organisms. The same will hold for other mental states. Why? The Functionalist answer is that mental states are functional states of the entire organism.

What does something have to be to qualify as a mousetrap? It just has to catch mice: The concept of "mousetrap" is functionally defined. On the Functionalist view, our concepts of mental states are functionally defined, too.

As a functional state, a mental state is triggered by a particular input that (a) results in particular behaviors, but (b) also triggers further mental states down the line.

Functionalism and multiple instantiation are often expressed using a metaphor of software and hardware, but the historical connection between Functionalism and computers is even closer. In the 1930s, Alan M. Turing

developed a general conception of computation: the Turing machine, thought of as operating step by step on individual symbols. How it functions at any step depends on its inner states, but the symbols also change its inner states. In introducing Functionalism, Hilary Putnam proposed that mental states are like the states of a Turing machine.

If mental states are functional states, they might be instantiable not only in living organisms but in machines. If so, the theory entails that an appropriately functioning robot would have beliefs, memories, and hopes, just as we do. If so, we could do psychology by doing robotics. Because of the change in perspective, Functionalism is ultimately noncommittal as to stuff. Although Functionalists tend to be Materialists, the position is, in fact, consistent with various views of what the universe is made of—including Dualism. ■

Suggested Reading

Lawrence Durrell, *The Alexandria Quartet*, *Justine*, and *Balthazar*.

Hilary Putnam, "The Nature of Mental States," *Art, Mind, and Religion.*

Ludwig Wittgenstein, "The Blue Book," *The Brown and Blue Books,* pp. 46–56.

Questions to Consider

Reflect on these two thought experiments:

1. You arrive as an explorer on another planet and encounter a new kind of organism that looks and moves roughly like a land-going squid. What would convince you that this creature is intelligent? What would convince you that this creature can feel pain?

2. An inventor has announced the first genuinely mechanical organisms. They look and move roughly like metal-jointed squids. What would convince you that they are intelligent? What would convince you that they can feel pain?

What Is It about Robots?

Lecture 7

People have long been fascinated with machines that act like people. One of the things I want to do in this lecture is to trace the history of that fascination. What is it that we see in robots? What makes them so fascinating?

What is so fascinating about robots? The concept of machines that are like people has an extensive history in myth and art, stretching from Homer's *Iliad* and the myth of Pygmalion in Ovid's *Metamorphoses* to such movies as *Blade Runner* and the *Terminator* series. Real robots also have an extensive history, from Descartes' influence on the golden age of automata in the 1600s and 1700s to their impressive use in various fields today. One purpose of this lecture is to explore our enduring fascination with robots by tracing their history in both art and reality.

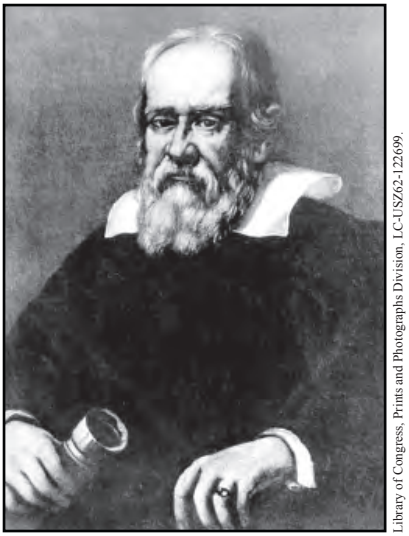
Robotic development has increased dramatically since the late 20th century. This lecture focuses in particular on contemporary robots, emphasizing both the promise they may offer and the threat they may pose. The inventor and theorist Ray Kurzweil foresees a "coming singularity," in which our machines become more intelligent than we are; roboticist Hans Moravec welcomes this possibility as the next step in evolution. The prospect of human-like robots also raises ethical issues of how we should treat our machines.

Robot is a recent term, but robot-like beings have a long history in myth, literature, and film. The term *robot* was introduced in Karel Capek's play *R.U.R. (Rossum's Universal Robots)*, first performed in 1921. Homer's *Iliad* has several references to robot-like beings. Romance and robots mix in the myth of Pygmalion from Ovid's *Metamorphoses*. Pygmalion creates a sculpture in the form of a woman and falls in love with it, and Venus brings the sculpture to life.

In the 1890s, Jean-Léon Gérôme painted the theme of Pygmalion twice. Ovid's myth inspired George Bernard Shaw's 1916 play, *Pygmalion*, which was the basis for the musical *My Fair Lady*. The Pygmalion theme appears

again in the movie *Blade Runner*. Myths of robot-like beings appear in other times and cultural contexts, including the medieval Jewish myths of the golem.

To recurrent themes of creation, power, protection, and love, the film robots of Hollywood add the theme of technology gone wrong. By the time the golem appears in film in 1915, he runs amok. In Fritz Lang's 1927 *Metropolis*, a female robot tries to destroy both people and machines. Technology gone wrong is a continuing theme in such movies as *Westworld*, *2001: A Space Odyssey*, and the first movies of the *Terminator* series. A more recent theme is sympathetic robot consciousness, evident in the later *Terminator* movies, *Blade Runner*, *Artificial Intelligence*, and of course R2-D2 and C-3PO of *Star Wars*.



Library of Congress, Prints and Photographs Division, LC-USZ62-122699.

Galileo's conviction by the Catholic Church served as a warning to Descartes, who hid his own work for fear of a similar fate.

The history of real robots is equally fascinating. What was the first "real robot"? That question reveals as much about our concept of robots as it reveals about the history of machines. Archytas of Tarentum, a friend of Plato, is reputed to have built a mechanical bird. Sometime between 100 B.C. and A.D. 100, Hero of Alexandria constructed a series of steam-driven religious altars, some of which incorporated moving statues. The 1354 Strasbourg clock incorporated several automata. In 1495, Leonardo da Vinci may have built a moving knight in armor.

Descartes' work inspired the golden age of automata. Lecture One mentioned the legend of Descartes' daughter, Francine. Whether that story is true or not, evidence exists that Descartes developed plans for automata. By 1633, Descartes had completed his work *De Mundo*. Because Galileo had been

convicted by the Catholic Church for similar arguments, Descartes decided to suppress the book. The view of animals and machines in *De Mundo* was developed in his later work. According to Descartes, animals are merely automata. They are purely mechanical and cannot feel.

Descartes' philosophical position inspired a burst of automata-building over several centuries. The centerpiece of that work was Jacques de Vaucanson's mechanical duck, mentioned in the work of de La Mettrie, Voltaire, and Goethe. The history of automata continues with increasingly sophisticated machines through the 18th and 19th centuries and all the way up to Walt Disney's theme-park attractions Mr. Lincoln and Pirates of the Caribbean. Some touted automata, including von Kempelen's chess-playing Turk, were fakes. All of these automata lacked genuine *autonomy*.

Humanoid robots are being developed in Japan to serve as nurses. Pet robots seem to offer therapeutic benefits.

Contemporary robots come in a variety of forms and have a variety of purposes. The first to be developed were industrial robots. More than a million are now in use, half of them in Japan. Robotic development is a major part of the space program. Spirit and Opportunity, robots designed to last 90 days on Mars, have continued to work for years. Surgery is increasingly being done by robots, with microbots in development that could swim through the human bloodstream.

Humanoid robots are being developed in Japan to serve as nurses. Pet robots seem to offer therapeutic benefits. Some experimentation has been conducted with "emotional" robots. Semi-autonomous robots are being developed for military purposes of surveillance, delivery, and aid, but the idea of attack by a robot army is chilling.

The future of robotics will bring a range of ethical dilemmas. What is the possibility of robots taking over? If we have something to fear, it will be at the point that we hand robot construction over to beings more intelligent than we are—our own robots. This situation is what inventor and theorist Ray Kurzweil means by the "coming singularity." Could we voluntarily limit the development of more intelligent robots? Would we? Should we? Roboticist

Hans Moravec sees a glorious future in this prospect. Our true descendants will be our "mind children"—robots that will replace us as the next stage in evolution.

Another set of ethical dilemmas concerns the robot as victim. If Functionalism is right, a machine could have real perception, emotion, pleasure, and pain. Wouldn't it also have ethical rights? We could design machines that want to do our dirty work, but what right do we have to design another being for our own ends? ■

Suggested Reading

Fritz Lang, *Metropolis*, 1927, video.

Terrell Miedaner, "The Soul of Martha, a Beast" and "The Soul of the Mark III Beast," *The Soul of Anna Klane.*

Ridley Scott, *Blade Runner*, 1982, video.

Gaby Wood, *Edison's Eve: A Magical History of the Quest for Mechanical Life*, chapter 1.

T.I.L. Productions, *Vaucanson and His Remarkable Automatons*, http://www.automates-anciens.com/english_version/automatons-music-boxes/vaucanson-automatons-androids.php (shows what may be photos of the remains of Vaucanson's duck).

Questions to Consider

1. Explore this thought experiment: You are offered the opportunity to be one of the first users of the Universal Household Robot, a human-sized robot that will be installed in your house and will follow general English commands to mow the yard, do the housework, balance the checkbook, and tend the kids. According to the company selling the robot, it may still have a few bugs that need to be worked out. The company also insists that you sign a waiver for any negative consequences that may result from use of the robot. Do you accept the offer? What would be the advantages? What could be possible dangers?

2. Ray Kurzweil has predicted that our machines will pass us in intelligence by the year 2040. How plausible do you think that prediction is? What do you think the consequences will be? Should we worry about this possibility? What might help protect us against negative consequences?

Body Image

Lecture 8

How is that body image formed? How does the mind produce a body? In exploring that question, I'll emphasize a perspective different from that in much of the history of philosophy of mind.

This lecture offers a philosophical examination of a range of psychological phenomena involving our perception of our own bodies. An image of our own body is laid out across the sensorimotor cortex, but the importance and plasticity of that body image are only now becoming clear.

People who have lost legs or arms often continue to feel them as *phantom limbs*, sometimes with intense pain. This phenomenon was long thought to be merely an effect of nerve stimulation in the remaining stump, but V. S. Ramachandran's work shows that phantom limbs are better understood and treated in terms of body image in the brain. The brain's body image also turns out to be amazingly "plastic," or adaptable. A violinist's development of technique is evident not only in performance but literally in his or her brain; areas of the brain are recruited and developed for particular motor skills and sensitivities.

How do we learn our bodies? In some ways, the body is a mental construct. Human infants take a significant amount of time to learn how their own bodies move. In some ways, the body is also a social construct. We have all incorporated cultural stereotypes of "normal bodies" and "perfect bodies," the dark side of which are eating disorders, such as anorexia and bulimia.

This lecture also emphasizes a perspective different from that in much of the history of philosophy of mind: the perspective of the embodied mind. A standard philosophical thought experiment is the brain in the vat. How do you know you are not a brain in a vat? That approach to minds assumes that they could be removed from "normal bodies" in normal interaction with the world. Antonio Damasio labels it "Descartes' error." The Functionalist mind, in contrast, is an "embodied" mind.

The somatosensory cortex registers sensation from different areas in the body and reveals much about the mind's body. The areas of your body that are more sensitive to touch occupy a greater proportion of the somatosensory cortex. With time and a few toothpicks, you and a friend can map the sensitivity of different body areas. A model of a person in which more sensitive areas are portrayed as larger is called a *sensory homunculus*. This sense of *homunculus* is different from that used in discussing the inner theater in Lecture Four.

Just as the area of brain tissue corresponds to the sensitivity of body area, the organization of brain tissue corresponds roughly to the organization of the body. The head, mouth, and pharynx occupy about half of the somatosensory cortex, right-side up in the lower portion. The rest of the body occupies the upper area, upside down, with the feet at the top of the head. Some areas that are not next to each other on the body, such as the feet and genitals, are next to each other on the brain's map. Neurologist V. S. Ramachandran asks whether this proximity explains foot fetishes.

The mind's body is not permanent: The plasticity of the brain allows for a takeover of one area for another function. Finger-sensitivity areas in the brains of people who use their fingers for fine coordination are larger than those in the brains of people who don't. This enlargement has been shown for both violinists and those who read Braille. The general idea of brain-area takeover was introduced in the discussion of Einstein's brain in the first lecture.

Some amazing cases show the degree to which brain function can be made up by parts of the brain not intended for those functions at all. Experiments with ferrets have shown that it is possible to rewire visual perception to the auditory cortex: Ferrets can learn to see with those parts of their brains wired for sound. Surgeons removed the half of three-year-old Jody Miller's brain that controlled the left side of her body. Within days, the brain compensated so that she could walk out of the hospital.

One of the most fascinating manifestations of brain plasticity and body image is the phenomenon of *phantom limbs*. Lord Nelson, hero of Trafalgar, lost his right arm in battle. He took the persistent sensation that the arm was still there as "direct evidence for the existence of the soul." The American

physician Silas Weir Mitchell coined the term *phantom limb* on the basis of repeated experiences of the phenomenon in amputees from the Civil War.

The explanation for phantom limbs was long thought to be stimulation of the nerves in the stump. Ramachandran has proposed an alternative theory, backed up with a range of case histories. Ramachandran proposes that phantom limbs occur when areas of the brain previously assigned to the missing limb are taken over. In one patient, Ramachandran has been able to show that sensations from parts of the face and shoulder are read as sensations from a missing hand.

How do you relieve pain in a limb that no longer exists? If phantom limbs are an aspect of body image, it should be possible in some cases to treat the pain that often accompanies phantom limbs by treating body image. Ramachandran has successfully used a mirror box in which a limb "seen" as the missing one can aid in relieving discomfort.

If phantom limbs are an aspect of body image, it should be possible in some cases to treat the pain that often accompanies phantom limbs by treating body image.

You can perform an experiment that creates something like the sensation of a phantom limb. You'll need a friend, a cardboard box open at both ends, and a rubber hand of the type available from novelty shops. Sit at a table opposite your friend and put your right hand into one side of the cardboard box. You can no longer see that hand, but your friend can see it and touch it from the other side. Now put the rubber hand next to the box where you can see it, as if it were your right hand lying there. Have your friend draw simultaneous and identical patterns on the rubber hand (which you can see) and your real right hand (which you cannot). For about half of those who try the experiment, the rubber hand starts to feel like their own. The idea that body image can expand to include inanimate objects—tools, for example—is also familiar from everyday experience.

How do we learn our bodies? A distinction can be drawn between your conscious impression of your body and your unconscious *body schema*. In

learning a skill, what is first intensely conscious can later submerge into body schema. The overwhelming consensus is that our sense of our bodies is learned in this way. Human babies take months to learn that those things waving in front of them are their arms, under their control. To a great extent, our sense of our own bodies must be learned because our bodies change over time.

Some evidence suggests that parts of our body image may be hardwired at birth. Within the first few days, newborn infants are capable of imitating facial gestures at better than a random rate, suggesting that they may have a default body image of their own faces. Cases have been reported in which people who have never had limbs experience the sensation of phantom limbs, suggesting a default image of limbs in the brain.

Professor Josh Bongard and his colleagues have constructed a robot that learns its body. The star robot first constructs and tests theories of its own body form in terms of feedback from its sensors. It then constructs and tests theories of how to move in a particular direction, with startling results.

Our bodies are, in some sense, mental constructs. As Ramachandran said, "Your whole body is a phantom limb." Our bodies are also, in some sense, social constructs. What should our bodies look like? Historically, images of the "perfect body" have varied greatly. The "beautiful body" is statistically out of reach for almost everyone. A dark side of the social construction of body image is evident in the eating disorders anorexia nervosa and bulimia. ■

Suggested Reading

Josh Bongard, Victor Zykov, and Hod Lipson, *Resilient Machines through Continuous Self-Modeling*, http://ccsl.mae.cornell.edu/research/selfmodels/ (the star robot in action).

V. S. Ramachandran and Sandra Blakeslee, *Phantoms in the Brain: Probing the Mysteries of the Human Mind*, chapters 1–3.

Scientific American Frontiers, *Changing Your Mind*, 1997 video.

Questions to Consider

1. Consider the following thoroughly speculative thought experiment: If it were possible to hook up a human brain to an animal's body—a cat's, say—do you think the brain could "learn" the cat's body? Are brains flexible enough to do that, or can a human brain handle only a human body?

2. Short of actually trying the Frankenstein-like experiment in question 1, what kind of evidence would convince you that the answer was yes—that a human brain could learn a cat's body? What kind of evidence would convince you that the answer was no?

Self-Identity and Other Minds

Lecture 9

Descartes thought there is one thing of which I can be absolutely certain: "I think, therefore I am." William James added that there's one thing about our conscious experience that we're never uncertain about; we're never uncertain whose conscious experience it is—it's *mine*, of course. But who is this "me"?

"I think, therefore I am." But what is it to be me? The mirror test for self-identification, first used by Charles Darwin, shows that chimpanzees have a reflective self-concept, but other primates do not. Children acquire a self-concept and pass the mirror test at about the age of two.

What is it to be the "same person" over time? The ship of Theseus is the classic example of problems of identity, but the issue comes in a variety of forms. What makes your body the same body over time? What makes you the same person? Thought experiments in terms of science fiction teletransporters highlight important problems for major theories of personal identity. The problem becomes real in consideration of puzzling experimental results with split-brain patients. Could it be that you have two seats of consciousness?

How do we know what other people feel and think? Here again, an ancient philosophical question—the problem of other minds—meets contemporary work in psychology and neuroscience. A recent philosophical theory emphasizes the role of simulation in our understanding of other minds—our ability to put ourselves in another's place. Results in the study of autism, brain scans, and the *mirror neurons* of macaque monkeys offer supporting evidence.

This lecture explores both our sense of ourselves and our sense of other people. Descartes said, "I think, therefore I am." William James noted, "I am never uncertain as to whether this consciousness is mine." Bertrand Russell asked, "[W]ho is this *me*? Who is this *I*?" This lecture explores our concept of ourselves, how far it extends in the animal kingdom, and some of the puzzling questions about self posed by split-brain cases. This lecture

also explores the other side of the coin: our concept of other people or other minds.

Do animals have a sense of self? Charles Darwin put mirrors in front of young orangutans to see whether they would realize they were looking at themselves. Since Gordon Gallup's work in the 1970s, recognition in a mirror has become the standard test for sense of self. If an animal sees a mark on the forehead of its reflection, will it reach for its own forehead to further explore the mark? Chimpanzees immediately reach to touch the marks. Other monkeys do not. Gorillas generally fail the test, with the exception of Penny Patterson's Koko. Human children pass the test at about the age of two. Elephants also seem to recognize themselves in a mirror.

What is our sense of self? What is it to be the "same person" over time? The metaphysical problem of identity over time appears in a number of forms. A classical version is the ship of Theseus. In Greek legend, Theseus slew the Minotaur and brought the youth of Athens home. Plutarch writes that the ship of Theseus was preserved for generations in honor of the feat. Plank by plank, each part of the ship is replaced. In his *De Corpore* of 1655, Thomas Hobbes adds a further spin: What if the old pieces are crutched together in the junkyard? Which is the real ship of Theseus, the pristine ship celebrated in the harbor or the assemblage of rotting timbers in the junkyard?

The problem of identity also appears in other forms. Given that your cells are constantly replaced, what makes your body the same body over time? You are a walking ship of Theseus. Hobbes took a pragmatic approach to questions of identity: The answer depends on what you are talking about. Conditions of identity for one kind of thing may be different than for another.

When is something the same person? In his *Essay Concerning Human Understanding* of 1689, John Locke gives a Functionalist account of when someone is the same man. In order for someone to be the same man, the same functional organization must be continuous over time. Locke gives a different account for when someone is the "same person." In this case, it is continuity of consciousness through memory that Locke thinks is important. Locke's view faces some problems. If the "same consciousness" were tied so

closely to memory, amnesia patients would automatically and invariably be different people than they were before they experienced amnesia.

Teletransporter thought experiments pose problems for many accounts of personal identity. The teletransporter maps all information about the chemical composition of your body and brain. The information is then sent to Alpha Centauri, where a perfect duplicate is assembled. "Ah," you say as you step from the teletransporter, "here I am." Is it really you that stepped out on the other side? Consider a "branching" case, in which the signal is sent to two different places, two people step from the transporters, and they go on to live two different lives. Both cannot be identical to you because they are not identical to each other. No one has used teletransporter examples better than the Oxford philosopher Derek Parfit.

In the end, bodily continuity, causal continuity, and continuity of memory do not seem to fare well as accounts of personal identity. Locke's identification of the "same person" with the "same consciousness" seems right but simply shifts the question: What is it that makes a consciousness mine? Just what is this sense of self?

Split-brain cases offer real instantiations of the philosophical problem. The corpus callosum is a thick connection of nerves that transfers information from one side of the brain to the other. When it is severed as a last-ditch effort to control severe epilepsy, lines of communication between the hemispheres of the brain are cut.

Roger Sperry's experiments show that under specific test conditions, the behavior of split-brain patients can be very strange. A woman shown a picture of a cup in her right visual field can answer what it is. When shown a picture of a spoon in her left visual field, she says she can see nothing, but when asked to find the object with her left hand, she does so successfully. According to Sperry, "Everything we have seen indicates that the surgery has left these people with two separate minds … that is, two separate spheres of consciousness." In one case, when asked what he wanted to be, a young split-brain patient answered "draftsman" with his left hemisphere but spelled out "automobile racer" with his right hemisphere.

How do we *know* that other people and consciousnesses exist? A standard answer to the *problem of other minds* is the *argument from analogy*: We know that other minds exist by inference or analogy from our own case.

Philosophers and psychologists have used the term *mind reading* to label our ability to read one another's mental states. How do we do that? One hypothesis, known as the *"theory" theory*, is that we have a theory that links behavior and context to mental states. Another hypothesis, known as the *simulation theory*, is that we know what someone else is feeling by directly simulating his or her situation in our minds.

Some examples can be interpreted in terms of either the "theory" theory or the simulation theory. In the *false belief* test, a child sees Katie put a marble in a drawer. When Katie is gone, the child sees Sally move Katie's marble. When Katie returns, where will she think the marble is? Before the age of three or so, children say that Katie will think the marble is where Sally moved it. After about the age of four, they say that Katie will think the marble is where she originally placed it.

Another hypothesis, known as the *simulation theory*, is that we know what someone else is feeling by directly simulating his or her situation in our minds.

Other examples favor the simulation theory. The majority of normal children and children with Down syndrome pass the false belief test by the age of four, but only a small minority of children with autism pass. Autistic children do well on many theory-like cognitive tests, while Down syndrome children do poorly on many. These results suggest that it is not a theory that is being mastered in mind reading. Results from brain scans show that the same areas of the brain are activated when a strong emotion is seen and when it is felt. This finding, too, fits the simulation theory. Studies with macaque monkeys have shown that *mirror neurons* are activated both when the monkey sees a particular action performed and when it performs that action—precisely what the simulation theory might predict. ■

Suggested Reading

Marc D. Hauser, *Wild Minds: What Animals Really Think*, chapter 5.

John Locke, *An Essay Concerning Human Understanding*, Book II, chapter XXVII.

Derek Parfit, *Reasons and Persons*, chapters 10–11.

Questions to Consider

1. Here's another thought experiment: Your family is waiting for you on Alpha Centauri, and the teletransporter crew is ready. In a few seconds, someone will step out of the teletransporter at the other end, and it will be you. Are you ready to step into the transporter, or do you have doubts? If you hesitate, what is it that you are worried about?

2. Review what you have learned in this lecture about mind reading, "theory" theory, and simulation theory. What evidence from your own experience favors one theory over the other?

Perception—What Do You Really See?

Lecture 10

This lecture and the next are structured around experiments from psychology and the neurosciences regarding perception. I want to put those experiments to work in guiding us through the philosophical debates regarding the nature of perception.

What do we really see? What do we really hear? The focus of this lecture is on the Empiricist theory of perception, which argues that what we really perceive are not things in the world but subjective sense-data—colored patches in our visual fields, for example. It is from these private sensations that we infer the existence of real things in an objective world. This lecture centers on several auditory illusions and experiments. It also describes visual experiments that seem to support the Empiricist picture.

Despite the theory's appeal, the lecture argues that Empiricism proves inadequate as a picture of perception. Core problems appear in the central concept of inference, in the reintroduction of the "little man" in the inner theater, and in the question of whether the theory really does justice to the data of experience.

Three very different theories have been put forth to explain perception: an Empiricist theory, an Intentionalist theory, and an Evolutionary theory. The focus of this lecture is on Empiricism. Intentionality and Evolution will be the focus of the next lecture.

What do we perceive? The Naïve Realist view says that we see things … things in the world. Naïve Realism takes perception to be a two-place relation. We have immediate perceptual contact with the things themselves. But contact seems to vary with different senses. Do you hear the locomotive or just the rumble of its engine? Do you smell the cookies or just the aroma of the cookies?

For seeing, touching, and perhaps tasting, perception seems to be a *two*-place relation. For the other senses, it is tempting to think of perception as a *three*-place relation. In smell, perception seems to be a relation among you, the aroma of the cookies, and the cookies themselves. In hearing, perception seems to be a relation among you, the sounds of the locomotive, and the locomotive itself. A little scientific knowledge pushes one to a three-place relation for all the senses. Sight is a three-place relation among you, the light reflected from an object, and the object itself.

Sight is a three-place relation among you, the light reflected from an object, and the object itself.

In an Empiricist theory, perception is even less immediate. The theory can be introduced in terms of illusions. Artists are familiar with the fact that things at a distance look bluer. If an orange is put under a green light, it looks distinctly gray. The Shepherd tones offer an auditory example. The bells seem to descend endlessly in pitch, but they couldn't possibly do so.

If you cannot tell the difference between something gray and something orange under a green light, what is it that you really see? According to the Empiricist theory, what you really see are sense-data, the colored patches in your field of vision, for example. Everything else—your knowledge of the external objects you are looking at, for example—is the result of an elaborate inference from immediate sense-data. Empiricist theory claims that all perception is a three-place relation among you, your sense-data, and the objects from which the data stem.

The theory appears with variations in the work of the Empiricists: John Locke, Bishop Berkeley, and David Hume, writing in the first half of the 1700s. Is Empiricism a psychological theory or a philosophical theory? It is both. Empiricism was a dominant theory well into the 20th century.

Empiricist theory has two important parts. The first part relates to the theory's fundamental entities: sense-data. The theory explains how things appear in terms of how other things are. When the orange looks gray, something really

is gray: your sense-data. The classic hallucination in Shakespeare's *Macbeth* fits the Empiricist theory exactly:

> Is this a dagger which I see before me,
>
> The handle toward my hand? Come, let me clutch thee.
>
> I have thee not, and yet I see thee still.

The second part of the theory is inference. What you directly perceive are sh=ense-data. From those, you infer the existence of external objects. Some psychological evidence seems to support the Empiricist picture. The *glissando illusion* was discovered by Diana Deutsch of the University of California at San Diego, one of the primary researchers in the psychology of sound and music. Although it's actually composed of little bits that are cut up, the glissando sounds continuous. The two horizontal lines in this picture are exactly the same length. Why does the top line seem longer than the bottom line?

The Empiricist answer is that you read the vertical lines as if they were railroad tracks going away from you into the distance. If they were,

something that appeared the way the top line does would have to be longer; thus, you see it that way.

In Diana Deutsch's *phantom words*, different people hear different words repeated: *respond*, *Congress*, *conscious*, *Christmas*, *miss me*, *mistress*. The sounds are actually syllables of *Boris* alternating in the two speakers. The process of filling in also seems to go along with the Empiricist picture of perception. This experiment shows you your blind spot:

Close your right eye and hold the page about a foot in front of you. Focus with your left eye on the x. Although you will not be looking at it, you will be able to see the dot to the side. Continuing to focus exclusively on the x, move the page slowly closer to your eye. When the page is about six inches from your eye, the dot will disappear. All you will see is white paper to the left of the x. If you keep focusing on the x and move the paper closer, the dot will reappear when it is about three inches from your eye. An experiment created by psychology professor Arthur Samuel shows that the same kind of filling in happens with what we hear.

How good is the Empiricist picture of perception? Despite its appeal, it has a number of problems. One problem is in the use of the term *inference*. Did you infer the continuity of the glissando, or did you just hear it that way? What is going on seems too simple and immediate to be inference. The term *inference* makes the process sound more cognitive and deliberative than anything we have evidence for. Doubting the use of the term *inference* does not challenge the data, but it does challenge the interpretation of the data.

The Empiricist picture seems to reintroduce the homunculus. Who is seeing the sense-data, and who is doing the inferring?

An Empiricist might say that the inference involved is unconscious inference or implicit inference, but this explanation creates another problem. If the inference is unconscious, the inference step in the theory is invisible. But the Empiricist insists that first we see sense-data, then we make an inference. If the inference is invisible, why not think it comes first? In many of our examples, that seems to be how it works. By the time you have sense-data, the processing has already been done.

The specific problems also introduce a more general reflection. The story that Empiricism tells about perception is a cognitive or rational story. Perception is portrayed as a rational inference from data. The history of philosophy is a history of people intensely devoted to rationality. It is perhaps not surprising that philosophers have tended to portray perception, too, as a rational process. ■

Suggested Reading

Michael Bach, *77 Optical Illusions and Visual Phenomena* http://www.michaelbach.de/ot/.

Diana Deutsch, http://psy.ucsd.edu/~ddeutsch (additional links with auditory illusions).

Donald D. Hoffman, *Visual Intelligence: How We Create What We See*, chapter 1.

Diego Uribe, *Truly Baffling Optical Illusions*.

Exploratorium: The Museum of Science, Art and Human Perception, http://www.exploratorium.edu/seeing/exhibits/index.html (a number of visual illusions).

Questions to Consider

1. When Grandma baked cookies for you, did you smell

- the cookies?
- the aroma of the cookies?

2. When the ambulance is approaching, do you hear

- its siren?
- the sound of its siren?

3. When the locomotive is coming down the tracks, do you see

- the locomotive?
- the sight of the locomotive?

4. If your answers to questions 1 through 3 were not all a's or all b's, what explains the difference?

Perception—Intentionality and Evolution

Lecture 11

In this lecture, I want to look at two very different approaches to perception, the Intentional and the Evolutionary approaches. You'll remember that according to the Empiricist theory, what we immediately see [are] not objects but sense-data. Our contact with the objects themselves is indirect. We infer the existence of outside objects on the basis of our sense-data.

What really happens when we see? Intentionalist and Evolutionary approaches offer alternatives to the Empiricist picture. In the Intentionalist picture, which can be traced back to the work of Franz Brentano in the 1800s, the "aboutness" of perception is essential to it. Brentano's slogan was: "All perception is perception of." The strange story of Oliver Sacks's patient who mistook his wife's head for his hat fits the theory nicely. So does a wide range of other work in the brain sciences on agnosia and prosopagnosia—the inability to recognize faces. But is it true that *all* perception is "perception of"?

The Evolutionary approach represents a different and more recent theory. What we should expect to find in perception, the Evolutionary theorist says, is not a single tidy picture but an assorted bag of adaptive tricks. Because we are evolved creatures, we will arrive with all the bits and pieces of perceptual equipment that proved evolutionarily successful in the past. Both the results of easy experiments and data from the neuroscience of perception are examined as evidence supporting the Evolutionary approach.

One objection to Empiricism was that the notion of inference seems too conscious, deliberate, and rational. That is precisely where the Intentionalist steps in. When we feel feathers, we do not feel sense-data from which we infer feathers. We feel the feathers directly. Content is not added; perception comes with content.

The theory traces to Franz Brentano, writing in the late 1800s. Bretano's slogan was: "All perception is perception of." The term *Intentionalist* comes

from a term used by the Scholastic philosophers of the Middle Ages to mean "conceptual content."

Intentionalism accords with a range of results from the brain sciences. *Agnosia* means "not knowing" and is used to classify a range of cases in which perceptual content seems to drop out. Oliver Sacks talks of a patient, Dr. P., who suffers from agnosia. Dr. P. can describe things before him in detail. "*What is this?*" "About six inches in length … a convoluted red form with a linear green attachment." But Dr. P. cannot identify the thing as a rose.

Prosopagnosia is an inability to recognize faces. To a person with prosopagnosia, faces are no more individuating than elbows. Some with prosopagnosia cannot even recognize their own faces. Often a person with prosopagnosia will not recognize family members or close friends until they speak—it is visual recognition in particular that is affected. Measures of emotional response may indicate that emotional recognition is in place even when visual recognition is not.

Is *all* perception "perception of"? All things considered, how good is the theory? The theory seems to be incomplete in an important respect. Smelling the soup is qualitatively different from tasting the soup. Seeing the train approach is qualitative different from hearing the train approach. These perceptions are "about" the same thing but are qualitatively different. An account in terms of "aboutness" seems to leave out that qualitative difference. But that qualitative character is what the account of the sense-data theorist is built on. Perhaps neither theory is adequate alone. Empiricism leaves out "aboutness," and Intentionalism leaves out qualitative character.

Further, we can identify apparent counterexamples to the theory. Suppose I just hear a bass rumble or see a pure blue field. Would not those be perceptions that were not perceptions "about" anything? The Intentionalist might say that these cases do have an object: These perceptions are "about" themselves. But that attempt to save the theory weakens it. Even the Empiricist's pure sense-data would count, and the Intentionalist's claim was that we do not perceive sense-data. The Intentionalist might say instead that the bass rumble and the blue field do not qualify as cases of perception. This response, too, weakens

the theory: It becomes not a theory of perception in general but merely of perceptual recognition.

Our ordinary talk of perception is complex. Some of the ways we think and talk about perception demand a concept of recognition and may, therefore, fit an Intentionalist theory. Other ways we think and talk about perception do not necessarily involve recognition. There, Intentionalism will fail.

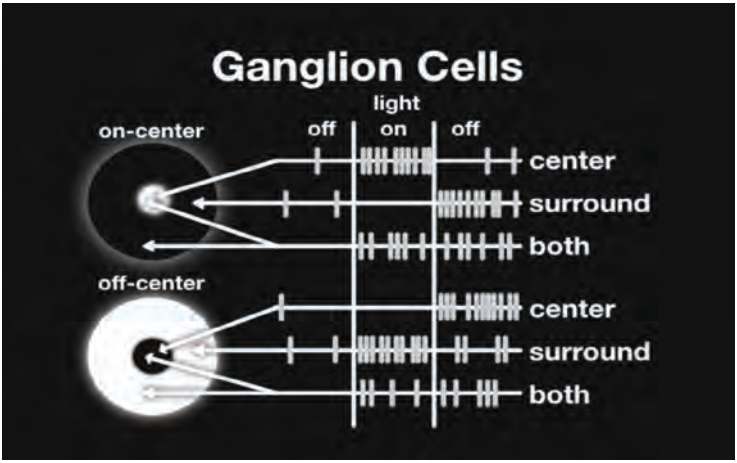
The Evolutionary approach is a more recent alternative to Empiricism. It offers a view of perception that is radically nonindividualistic and nonrationalistic. According to the Evolutionary approach, we are evolved organisms and can expect to carry a legacy of bits and pieces of perceptual equipment that proved evolutionarily successful in the past.

If you track a tennis ball moving in front of you, your eyes move in a smooth pattern. If you try to trace the same route without a moving ball, your eyes move in little jumps called *saccades*.

Saccades offer one piece of evidence in support of an Evolutionary approach. If you track a tennis ball moving in front of you, your eyes move in a smooth pattern. If you try to trace the same route without a moving ball, your eyes move in little jumps called *saccades*. You seem to have two systems of eye-tracking. Why? An Evolutionary explanation is that you have two different systems for two very different tasks: (a) tracking moving prey or predators and (b) scanning a space in front of you.

The neurological details of perception offer further support. Ganglion cells respond to *on-center* and *off-center* features in their receptive fields in the retina.

On-center cells fire when you shine a light just in the center of the receptive field or if light around the edges goes off. Off-center cells fire when a central light goes off or when light around the edges comes on. The neurons in the primary visual cortex V1 appear to be *line detectors* and *edge detectors*. From there, processing goes to at least 20 higher layers. Visual



processing is not deferred to some inner theater in the brain. This, too, fits an Evolutionary picture.

The phenomenon of *blindsight* offers further support. People who have suffered damage to the V1 area are cognitively blind. If you ask them to guess simple shapes in front of them, however, their answers are fairly accurate. Nicholas Humphrey, who first studied blindsight in monkeys, thinks that we see in two ways. One is conscious, through V1. When that is knocked out, an evolutionarily older form of vision can remain in place.

Which theory of perception wins? Paul Bach-y-Rita has developed an aid for the blind that uses a thin plastic film placed on the tongue. Electrodes on the tongue encode signals from a video camera. Within a short period, blind subjects can recognize shapes and motions well enough to catch a ball thrown to them. The part of the brain recruited for the task is the visual cortex, the same part you use to see.

In this lecture we have considered a variety of extraordinary cases: Dr. P's agnosia, blindsight, and blind subjects equipped with a device that stimulates their tongues. Which of these is really perception? Can these people see or not? The major conceptual mistake may be to think that such questions have a single answer. Important conceptual work needs to be done in distinguishing

and articulating subtleties of the concepts of seeing, recognizing, reacting to, discerning, observing, watching, looking at, attending to, seeing that, glancing, and the like. ■

Suggested Reading

Michael Bach, *77 Optical Illusions and Visual Phenomena*, http://www.michaelbach.de/ot/.

Francis Crick, *The Astonishing Hypothesis: The Scientific Search for the Soul*, chapter 10.

Oliver Sacks, *The Man Who Mistook His Wife for a Hat*. chapter 1.

Exploratorium: The Museum of Science, Art and Human Perception, http://www.exploratorium.edu/seeing/exhibits/index.html (a number of visual illusions).

Questions to Consider

1. In this lecture series, we have often tried to decide not whether a theory was absolutely right or absolutely wrong but what might be right about it and what might not be.

2. The Empiricist says that perception is a two-step process: (1) Sense-data are received and (2) inference is then made on the basis of the sense-data. When you see a house, a hammer, or a friend's face, what you see is really the result of an elaborate process of inference from sense-data.

3. The Intentionalist says that perception is more immediate than that. "All perception is perception of." Content is not something that needs to be added by inference. Perception *comes* with content: What you see is a friend's face.

4. To what extent do you think each theorist is right? To what extent is each wrong? To what extent might they be talking past each other?

A Mind in the World

Lecture 12

Just as we go wrong if we think of the mind in isolation from the organism, we can also go wrong if we think of the organism in isolation. Organisms are what they are, they do what they do, they function as they function largely because of the environment of which they're a part. Our understanding of minds demands that we understand both the organisms of which they form a part and the environments crucial to those organisms.

This lecture explores the "mind-in-the-world" approach as it has been developed in psychology, philosophy, and robotics. J. J. Gibson's psychological theory of "affordances" is outlined against the background of its development in training World War II pilots. The classic psychological experiment of inverting lenses directly addresses questions of the mind in the world and is considered in terms of Alva Noë's three-stage analysis of the experiment. Philosophers Andy Clark and David Chalmers offer a thought experiment intended to show that memories, beliefs, and thinking itself can consist of parts of the external world outside our skins. This lecture examines each of these views, highlighting both core truths and misleading overstatements. Rodney Brooks has arrived at a similar theory in robotics by intentionally negating the assumptions of others in the field. Brooks's robots are built with an eye to embodied intelligence: a mind in the world rather than separate from it.

The central idea accords with both Functionalist and Evolutionary approaches. The Functionalist claims that mental states are functional states of an organism, and organisms function in environments. For the Functionalist, the link from mind to world is direct.

The young Charles Darwin was fascinated by differences in the Galápagos finches. Why are there so many different kinds of finches? Darwin concluded that so many different kinds of finches exist because so many different environments exist. Environmental pressure is a crucial element

in the evolutionary process. It is environments that do the selecting in natural selection.

J. J. Gibson argued that the mind can be understood only in terms of the world of which it's a part. Gibson first developed the theory in training World War II pilots. The core of Gibson's theory is that what we perceive are not sense-data. We do not see base-level sensations of any kind from which we have to infer objects and their motions. According to Gibson, "Eyes evolved so as to see the world, not a picture." What we perceive, Gibson says, are "affordances." Affordances are possibilities for action. Perception is direct, rather than inferential. It is tied directly to meaningful action in the world. Gibson's theory is tied both to the Evolutionary approach and the Intentionalist theory.

Gibson's theory has led to important research, but it is not adequate if taken as a complete theory of perception. Does it tell us the full story of what is happening in perception? We are sometimes fooled: We think we see an open door, but it turns out to be a clever painting. Why? Affordances are defined as real features of the environment—for example, real open doors. For that reason, Gibson's theory is unable to explain perceptual error.

Is it true that affordances are all we ever perceive? In order to explain illusions, we could vary the theory to say that what we perceive are apparent affordances: apparently open doors, for example. What do the real and apparent affordances have in common? The answer seems to demand some lower-level perceptual elements. But that explanation violates the anti–sense-data spirit of the entire theory. We might still agree that we cannot understand the mind unless we understand how the mind works in the world. But we need a critical qualification to Gibson's theory: that "perceiving affordances" is not the whole story of perception.

The classic psychological experiment of the inverted lenses underscores the theme. What happens if you put on glasses that reverse things right to left? The philosopher Alva Noë has analyzed the result in terms of three stages:

- The first stage is one of disorientation.

- In the second stage, things become clear but are distinctly reversed.

- In the third stage, the world "rights itself" again.

The third stage offers clear support for the "mind-in-the-world" approach but remains philosophically perplexing. In a high school dramatization, the experience of inversion was illustrated by turning the camera over. The subjective experience of things coming "right again" was illustrated by turning the camera upright again. An alternative view is that one simply gets used to the inversion and is once again able to navigate in the world. The deepest philosophical question is whether these two accounts are really different. Perhaps they amount to the same thing.

Philosophers Andy Clark and David Chalmers have pressed the idea further in terms of what they call "extended mind." Where does the mind stop and the rest of the world begin? Clark and Chalmers argue that a range of our mental activities can extend beyond our skin, composed in part of things in the world. They offer the thought experiment of Inga and Otto. Inga has a background belief that the museum is on 53rd Street. Otto suffers from Alzheimer's disease but carries a notebook in which he writes down information. He checks his notebook and finds that the museum is on 53rd Street. Inga's beliefs and memories are in her head in this example. Otto's beliefs and memories are in his notebook.

If that account were right, I could announce a complete and total cure for *anterograde amnesia*, a condition in which one is unable to lay down new long-term memories. The cure: a pencil and notebook. Clark and Chalmers concede that Otto's beliefs or memories would not normally be classified as such, but the two philosophers think they should be.

Many of the things we do involve parts of the external world in crucial ways. The video game Tetris allows you to think by rotating shapes physically. The calculations involved in balancing a checkbook would be much more difficult without pencil and paper. A key to logic puzzles in standardized tests is to make a sketch or diagram. The mathematician and philosopher Alfred North Whitehead made a similar point in terms of looking for the right symbolism

to use in approaching a problem. Clark and Chalmers mention language and linguistic interaction with other people as a form of "extended mind."

Innovative work in robotics has shown a parallel emphasis on the mind in the world. Rodney Brooks is head of the Robotics Lab at MIT and founder of the iRobot Company. In autobiographical notes, he has said that he tries to find an assumption that everyone is making, then negates that assumption. One goal for robotics has long been the creation of humanoid robots—robots that walk, talk, and think like people.

One goal for robotics has long been the creation of humanoid robots—robots that walk, talk, and think like people.

Brooks points out that evolution did not start with people, and he thinks that robotics should not either. Brooks attempts to build systems bit by bit, moving slowly toward higher intelligence, just as evolution did.

NASA uses robots for space exploration and approached Brooks for ideas in designing a 100-pound robot appropriate for planetary exploration. Why one 100-pound robot, rather than 100 one-pound robots? Brooks's proposal was for a crowd of small robots, "fast, cheap, and out of control." In developing robots, Brooks rejects the idea that they must form representations in a central processor. Why not use distributed intelligence? Throughout Brooks's work runs the idea of embodied intelligence: a mind in the world rather than separate from it. ■

Suggested Reading

Rodney Brooks, "Intelligence without Representation," *Artificial Intelligence* 47.

Andy Clark and David Chalmers, "The Extended Mind," *Analysis* 58.

J. J. Gibson, "Autobiography," in *Reasons for Realism: Selected Essays of James. J. Gibson*, Edward Reed and Rebecca Jones, eds.

Errol Morris, *Fast, Cheap, and Out of Control*, 1997, video.

Questions to Consider

1. If Clark's and Chalmers's extended-mind theory is right, part of your memory is not in your brain but, rather, in a file drawer or a photo album. To what extent does that fit with your experience? To what extent does it not?

2. Consider this thought experiment: NASA has asked you to plan a robot exploration of the planet Jupiter. The agency can afford a payload of only 100 pounds of robotics and must decide among (a) a single large robot, (b) several medium-sized robots with somewhat reduced capabilities, or (c) Rodney Brooks's recommendation of 100 very small robots, each with more reduced capabilities. What strategy do you recommend? Why?

A History of Smart Machines

Lecture 13

One of our greatest assets is our ability to think. Logical thinking and mathematical calculation in particular are essential and powerful human abilities. But as problems become complex, logical thinking and calculation can require a great deal of tedious effort. Throughout history, humans have tried to make it easier.

This lecture traces fascinating stories of computing machines, from the Antikythera machine of 100 B.C., to legends of mechanical calculating heads in the Middle Ages, to Charles Babbage's designs for steam-driven computers in the 1840s. Philosophers have played a major role in the development of smart machines. The crucial concept of logic begins with Aristotle. Both Pascal and Leibniz built early mechanical calculating machines. The foundations on which all contemporary computers operate can be traced to developments in logic in the 1800s and to Bertrand Russell and Alfred North Whitehead's *Principia Mathematica*.

The story begins with the logic of the ancient Greeks and culminates in the computers of today. What do logic and computers have in common? Everything. The subject of this course is not logic, but the fundamental concepts of this discipline will prove useful. The history of logic can be traced to the towering figure of Aristotle. Primary among Aristotle's logical works are the *Categories*, the *Prior Analytics*, the *Posterior Analytics*, and *De Interpretatione*.

The fundamental goal of Aristotle's logic was to capture, to systematize, and to formalize validity. The validity of an argument is its logical strength: whether the conclusion really follows from the premises. The fact that the conclusion follows from the premises does not tell us whether the premises are true. Validity does not guarantee truth, but it does guarantee truth preservation: *If* the premises are true, the conclusion *must* be true.

Aristotle's fundamental insight, echoing through the entire history of logic, is that validity is a matter of the structure of an argument. Different arguments with identical structure will have identical validity or invalidity:

Argument 1	Argument 2
Many medicines require a prescription.	Many physicists are Reductive Materialists.
Everything that requires a prescription requires a doctor's signature.	Everyone who is a Reductive Materialist is a Monist.
It follows that many medicines require a doctor's signature.	It follows that many physicists are Monists.

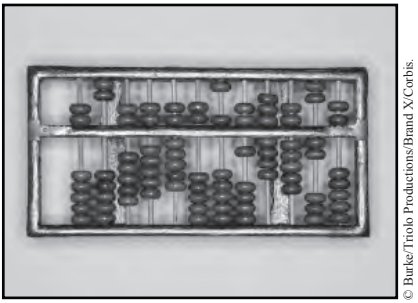
Specific content does not matter to structure or to validity. We should, therefore, be able to detect validity in terms of some stripped-down representation of the form of the argument:

- Many Xs are Ys.
- Everything that is a Y is a Z.
- It follows that many Xs are Zs.

Perhaps that aspect of thinking can be mechanized.

What was the first computer? Trying to answer that question tells us as much about our concept of computers as it does about the real history of real machines. It is tempting to count the abacus as the first calculating machine. But it can also be seen as merely a way of keeping a running record of human calculations: as a form of notation. Is Stonehenge an eclipse-predicting device from 2500 B.C.? The Antikythera machine was a clocklike mechanism from the 1st century B.C. capable of predicting the motion of the Moon and planets. Machines of medieval legend include a mechanized abacus and a brass calculating head in which figures appeared as teeth. History has seen the steady development of calculating devices for ballistics and navigational instruments.

Addition is easy, but multiplication is hard. The history of arithmetical calculating machines was driven largely by that arithmetical fact. John Napier of Scotland introduced two ways of handling multiplication in terms of addition. One of his inventions was Napier's bones, the first pocket calculator. Napier's bones was a set of ivory sticks marked with numbers that could be used to multiply merely by adding. Napier also invented logarithms, which allow the multiplication of two numbers by adding their exponents. The slide rule uses the principle of logarithms to mechanize calculation. Slide rules were used as late as the 1960s.



It is tempting to call the abacus the first calculator, but actually it only keeps track of calculations.

A series of machines in the 1600s began to look more like computing machines. In 1623, after a visit with Kepler, William Schickard produced a machine capable of multiplication. The philosopher and mathematician Blaise Pascal followed with the Pascaline. Leibniz produced a multiplying machine in 1672.

Previous machines are overshadowed by Charles Babbage's designs for steam-driven computers. Babbage took great delight in finding errors in published logarithm tables. He designed a machine to calculate logarithms and print them mechanically. Babbage's Difference Engine exhausted Parliament's financial patience. It was never completed.

In the face of disappointment, Babbage developed a grander plan for the Analytical Engine. It would have been able to perform and combine all arithmetical operations. The design is amazingly contemporary, with a "mill' and a "store" corresponding to the central processor and memory of contemporary computers. Processing was to be carried out on punched cards, invented for the Jacquard loom in 1801.

Another astounding character in the story of Babbage is Ada Lovelace, the only legitimate daughter of Lord Byron. She fell in love with the Analytical Engine, and some of her ideas for repeated loops and subroutines were incorporated in the final design. The Analytical Engine was never funded.

Contemporary computers are built even more directly from logic. Fifty years before Babbage, Kant had announced that logic was a completed science. There was no more logic to do. Kant could not have been more wrong. Logic was reborn in the 1800s through the work of George Boole and Gottlob Frege. The goal was the same as it had been for Aristotle: to systematize, formalize, and mechanize thought.

A major step in the development of logic and the birth of computers as we know them was Bertrand Russell and Alfred North Whitehead's *Principia Mathematica.* The purpose of *Principia* was to prove that all mathematics was essentially logic. To achieve that goal, Russell and Whitehead showed that a few simple logical symbols—AND, OR, and NOT, for example—were enough to give all the numbers, functions, operations, and transformations of mathematics.

***Principia Mathematica* was intended merely to settle an arcane philosophical dispute. It turned out to be the founding document for all contemporary computing.**

If that is true, we should be able to build a computing machine out of those simple components as well. The logical connectives can, in fact, be reduced to just one: NAND, meaning "not both …" With a little ingenuity, NANDs can be compounded to produce all the rest. Deep down inside, all digital computers work with the single logical connective NAND, operating on the 1s and 0s of binary code. *Principia Mathematica* was intended merely to settle an arcane philosophical dispute. It turned out to be the founding document for all contemporary computing.

Two major developments came after *Principia Mathematica*. One was the attempt to develop artificial intelligence. Here, Alan M. Turing is a major figure. Turing's work will be discussed in Lecture Fifteen. The other development shows important limits to logic. The paper by Kurt

Gödel that radically altered our view of mathematics was called "On Formally Undecidable Propositions of *Principia Mathematica* and related systems." Gödel's work will be discussed in later lectures on the topic of consciousness. ■

Suggested Reading

Aristotle, *De Interpretatione*.

Michael R. Williams, *A History of Computing Technology.*

Computer History Museum, http://www.computerhistory.org/about_us.html.

Science Museum, http://www.sciencemuseum.org.uk/visitmuseum/galleries/computing/ondisplay.aspx.

Questions to Consider

1. How do you think history would have been different if Babbage had succeeded—if computers driven by steam had been available before the Civil War? In what ways would history have been the same?

2. We have constructed strong machines to relieve ourselves of physical labor and to allow us to do things physically that we could not do otherwise. We have constructed smart machines to relieve ourselves of the labor of certain types of thought and to allow us to do things in the realm of thought that we couldn't do otherwise. Are there important differences between having machines do our mental work and having them do our physical work?

Intelligence and IQ

Lecture 14

The main question I'll address in this lecture is the dispute over whether it's one thing at all. Should we think of intelligence as one thing, in the singular, or should we think of intelligences—lots of them of different kinds—in the plural? Pursuing that question will take us through the history of IQ testing and a look at some alternative approaches.

What is this thing *intelligence*? Should we think of intelligence as one thing, in the singular? Or should we think of multiple intelligences, in the plural? This lecture examines that dispute. The history of the attempt to measure the *one* thing that is intelligence is the history of IQ testing. This lecture traces that history, from the measurement of skulls, through Alfred Binet's first IQ tests, to the adoption and misapplication of IQ testing in America. Are IQ and intelligence a result of genetics or environment? What do racial differences in IQ mean? These questions are examined with particular attention to the ways they were used in arguments for eugenics and social engineering.

Cases of prodigies and savants are considered as part of the evidence for an alternative approach in terms of *multiple* intelligences. Howard Gardner's work is examined in terms of its educational implications. The lecture closes with a first look at differences between natural and artificial intelligence: the chess competitions between grandmaster Gary Kasparov and IBM's Deep Blue.

The history of attempting to measure the *one* thing that is intelligence is the history of IQ testing. The history actually starts with *craniometry*, the attempt to measure intelligence by measuring the size of brains. Between 1821 and 1851, Samuel George Morton amassed more than 600 human skulls. When he measured cranial capacity, Morton found Caucasians to have the biggest brains, followed by Asians, Native Americans, and those of African descent. When Stephen Jay Gould reexamined the skulls 130 years later, he found that Morton's results depended on how he drew borderlines between groups and the proportion of male and female skulls in each group.

By the end of the 1800s, scientists were measuring the skulls of living people using the *cranial index*. One of those scientists was Alfred Binet, who measured the skulls of French schoolchildren. Binet became frustrated with his results. The poorer students sometimes measured "more intelligent" by the cranial index. When he found that his assistant got different results and that those results then influenced his own measurements, Binet gave up the cranial index in disgust.

When commissioned to identify schoolchildren most in need of help, Binet used an intentional hodgepodge of behavioral tests. Binet calculated results in terms of the mental age and chronological age of the child. German psychologist W. L. Stern suggested dividing mental age by chronological age, and the intelligence quotient, or IQ, was born. Binet emphasized that the test was not intended to measure some single thing called "intelligence." He warned that it should not be used as a basis for ranking normal children. He insisted that whatever the test measured, it should not be assumed to be innate or unchangeable.

In *Buck v. Bell* (1927), the Supreme Court upheld a Virginia law that called for sterilization of the mentally inferior. Although many states have since rescinded similar laws, the decision in *Buck v. Bell* has never been officially overturned.

Early in the 20th century, Binet's test was brought to America and renamed the Stanford-Binet test. When the test was brought to America, all of Binet's warnings were ignored. The test was used for ranking people in general; it was treated as measuring a single thing called intelligence; and it was assumed that what it measured was innate. H. H. Goddard, who brought the test to America, labeled those who scored in the three lowest categories on the test *idiots*, *imbeciles*, and *morons*.

In *Buck v. Bell* (1927), the Supreme Court upheld a Virginia law that called for sterilization of the mentally inferior. Although many states have since rescinded similar laws, the decision in *Buck v. Bell* has never been officially overturned.

During World War I, IQ tests were used for officer selection. Test conditions were poor and uneven, and the results were influential in passing the Immigration Restriction Act of 1924.

Does IQ correlate with race? The average IQ of American whites is approximately 15 points higher than the average for American blacks, though the gap appears to be narrowing. The average for Hispanics falls in between. The average for Asian-Americans is 3 points higher than for American whites. Such data played a role in Arthur Jensen's work on race and intelligence in the late 1960s and in Richard Herrnstein and Charles Murray's *The Bell Curve: The Reshaping of American Life by Difference in Intelligence.* In all cases, we are dealing with averages: Individual variation within each group is far wider. Several of the prominent figures mentioned here later recanted the conclusions they drew from the data, including H. H. Goddard. Differences in scores, whether individual or in group averages, tell us nothing about causes. That information must be found elsewhere.

Is IQ a matter of one's genes or one's environment? To what extent can IQ scores be accounted for in terms of genetics and to what extent in terms of environment? The consensus is that genetic factors explain somewhere between 40 and 80 percent of IQ variability, with environmental factors accounting for between 20 and 60. Many researchers settle on 50/50. Studies of twins and siblings are most useful in trying to obtain an estimate. Although we know that both heredity and environment play a role in IQ, we do not know precisely how.

In order to avoid past mistakes, it is important to examine the argument that led from IQ to eugenics. The argument proceeds in three steps. The first step is the claim that IQ is hereditary. The second step is the notion that if IQ is hereditary, it must be immutable. The third step draws social conclusions from the first two. The argument has been used to support the claim that social inequities are unavoidable, that education programs for certain classes of people are a waste of social resources, or that we owe it to society to breed some kinds of people and not others.

Is IQ hereditary? The answer is a qualified yes. But the second step does not follow. Immutability does not follow from the fact that something is

hereditary. Height has a very high heritability, but average heights worldwide have increased dramatically over the last century. IQ scores worldwide have also increased.

The most dangerous step is the third one: the jump to social conclusions. Proponents claimed that the scientific data tell us we should breed some people and not others. No argument to that effect can be purely scientific. At that point, any argument must bring in appeals to ethics and the kind of society we want ours to be. These are questions of value, above and beyond what scientific data of any kind can tell us.

The history of IQ testing is the history of attempts to measure the *one* thing that is intelligence. The idea of multiple intelligences is an alternative approach. Contemporary IQ tests follow Binet's in using a variety of subtests. The Wechsler Adult Intelligence Scale includes the following:

- Picture completion.
- Digit span.
- Narrative.
- Vocabulary.
- Similarities.
- Object assembly.

One way of scoring the results is in terms of a single intelligence parameter. Another, pioneered by Louis L. Thurstone, is in terms of several primary mental abilities. Nothing in the data dictates one approach rather than the other.

One piece of evidence in favor of multiple intelligences is the phenomenon of prodigies.

- John Stuart Mill learned ancient Greek by the age of three. By eight, he had read Herodotus, Plato, and others in the original Greek.

- Mozart began to play the harpsichord at the age of three and was composing by the age of six.

- Steve Wozniak was developing sophisticated electronics while still in the fifth grade.

Library of Congress, Prints and Photographs Division.

Wolfgang Amadeus Mozart.

- The fact that prodigies excel in very particular areas is an argument in favor of specialized intelligences.

Cases of savants offer another piece of evidence.

- Kim Peek was the inspiration for *Rain Man*. Given the name of a small town, he can give the zip code and the area code and perform a mental MapQuest search.

- Peek is also a calendar calculator; he can give the day of the week of any date in history.

- Oliver Sacks recounts the numerical ability of a pair of twins who would trade primes.

- The fact that savants may test as unintelligent in other areas is another argument for the existence of distinct types of intelligence.

Harvard psychologist Howard Gardner developed a theoretical account of multiple intelligences. He used a number of sources to identify separate forms of intelligence: studies of prodigies and savants, cases of brain damage, distinctive developmental histories, abilities developed in some cultures but not others, and transference of training. Gardner proposes seven intelligences, with an individual representing each:

- Linguistic intelligence—T. S. Eliot.
- Logical and mathematical ability—Albert Einstein.
- Spatial intelligence—South Sea islanders or Picasso.
- Musical intelligence—Igor Stravinsky.
- Bodily-kinesthetic intelligence—Martha Graham.
- Interpersonal intelligence—Sigmund Freud.
- Intrapersonal intelligence—Gandhi.

In the 1990s, Gardner added naturalist intelligence, exemplified by Charles Darwin. Some have proposed adding emotional intelligence. The theory has clear educational implications. Gardner's vision is of an educational system that recognizes and cultivates multiple intelligences.

We can also compare human and artificial intelligence. In 1996, IBM's Deep Blue was the first computer to win a chess game against a reigning world champion. In 1997, Deeper Blue won the match. Does that show that Deeper Blue was smarter? The system could evaluate 200 million positions per second. Perhaps the amazing thing is not that Deep Blue could match human play but that a human with far inferior calculation and memory could match *it*. Human chess play seems to involve a different kind of intelligence: pragmatic pattern recognition. ■

Suggested Reading

Howard Gardner, *Multiple Intelligences: The Theory in Practice*, chapters 1–2.

Stephen Jay Gould, *The Mismeasure of Man*, chapter 5.

Questions to Consider

1. Perhaps Howard Gardner and others are correct and intelligence is many things rather than one. What difference should that make for our educational system?

2. What qualities would you most want to give a child? If you were to put the following traits in order of importance, what would you list first? Would you add any other qualities?

- Creativity.
- Ambition.
- Physical beauty.
- Intelligence.
- Religious sense.
- Tenacity.
- Appreciation of nature.
- Sense of humor.

3. Considering each quality separately, what environmental situations or tasks encourage its development?

Artificial Intelligence

Lecture 15

In this lecture, I want to introduce the history of artificial intelligence, or AI for short. There was one man crucial to the entire attempt, crucial in different ways at different times. He was as historically important for contemporary artificial intelligence as Einstein was for contemporary physics. That man was Alan M. Turing, a British mathematician and logician.

Could a machine be genuinely intelligent? Alan Turing proposed a test: Could you tell, on the basis of answers to your questions received on a computer monitor, whether you were communicating with a person or a machine? A machine indistinguishable from a person in such a test, Turing proposed, would be a genuinely thinking machine, intelligent in the full sense of the term.

Turing also predicted that we would have thinking machines fully comparable to people by the year 2000. What has happened to the dream of artificial intelligence since the invention of the Turing test? This lecture traces the history of a range of attempts, including both the embarrassing failures of "Good Old-Fashioned Artificial Intelligence" and the intriguing promise of Connectionism and neural nets.

During World War II, Turing was instrumental in cracking the Nazi code. The German Enigma machine operated with moving rotors, which effectively changed the substitution code with each keystroke. Surveying the billions of billions of possible combinations the Enigma machine could produce would take millions of years. The Nazis changed the codes each *day*. Part of the Allied strategy in cracking the code relied on predictabilities in messages. Part of the strategy relied on the fact that the Enigma machine could never encode a letter as itself. A major part of cracking the code was the construction of another machine that could sort likely possibilities with immense speed. Before the war, Turing had developed the Turing machine, an abstract model of the concept of computation in general.

In 1950, he published "Computing Machinery and Intelligence" in the philosophical journal *Mind*. That article gave us the Turing test. Turing outlines the Turing test in terms of an analogous party game. In the imitation game, the question is whether you can differentiate a woman from a man pretending to be a woman when both are behind a door and providing written answers to written questions. In the Turing test, the question is whether you can differentiate a person from a machine simply on the basis of answers on a monitor to questions asked on a keyboard. Can a machine think? Turing thinks the closest we can get to an answer to that kind of question is: Can a machine be built to give responses that are indistinguishable from those of a thinking person?

In the Turing test, the question is whether you can differentiate a person from a machine simply on the basis of answers on a monitor to questions asked on a keyboard. Can a machine think?

In 1950, Turing made a prediction: "I believe that in about fifty years' time, it will be possible to programme computers … to … play the imitation game so well that an average interrogator will not have more than 70 percent chance of making the right identification in five minutes of questioning." A version of the Turing test is run each year. Contestants submit programs in the attempt to win the $100,000 Loebner Prize.

Turing's prediction has not been fulfilled. Did something go wrong? The next major step in the AI story was the Dartmouth Artificial Intelligence Conference of 1956, which brought together many of those who would be instrumental in the development of the field. The participants at the Dartmouth Conference had two different goals and favored two different approaches. The goal of some of those who attended the conference was to understand human intelligence. The goal of others who attended was to build machines with new capabilities. Some thought the core of intelligence was symbol-processing. This approach came to be called *GOFAI*, meaning "Good Old-Fashioned Artificial Intelligence." Others thought we should build machines that operated on the basic principles of the brain, an approach

called *Connectionism*. The history of AI since has often been a history of competition between GOFAI and Connectionism.

The first successes were from the symbol manipulation approach. Allen Newell and Herbert Simon produced the Logic Theorist, which proved logical theorems in the style of Bertrand Russell and Alfred North Whitehead's *Principia Mathematica*. Marvin Minsky, founder of the AI laboratory at MIT, produced a machine to prove theorems in geometry. The future of AI seemed unlimited, and wildly optimistic predictions were made, none of which was realized. It became clear that many human abilities and much of human intelligence depend on something other than symbol processing, such faculties as pattern recognition, contextual sensitivity, and rough-and-ready categorizations.

The second approach advocated at the Dartmouth Conference was Connectionism. The terms *neural networks* and *parallel computing* are also used in reference to this approach. Anticipations of the idea of associative learning appear in the work of David Hume in the 1700s and John Stuart Mill in the 1800s, but the first working models are those of Warren McCulloch and Walter Pitts in the 1940s.

In 1962, Frank Rosenblatt developed the *perceptron*, a feed-forward neural network consisting of two connected layers: inputs and outputs. For any pattern a perceptron could instantiate, Rosenblatt's learning rule could train it to that pattern. Marvin Minsky and Seymour Papert attacked neural nets, showing that perceptrons could not instantiate certain patterns, such as *exclusive or*. Minsky and Papert convinced the field that Connectionism could not succeed. Grants for research on neural nets dried up.

Neural nets were reborn in the mid-1980s through the work of David Rumelhart, Jay McClelland, and the Parallel Distributed Processing Research Group. A new learning rule, *backpropagation of errors*, allowed the training of multilayer nets. Neural nets are good at pattern recognition, the kind of task that stumped GOFAI. Garrison Cottrell's face-recognition net offers a good example.

For one goal of artificial intelligence, neural nets can be frustrating. Opening up a neural net after training may not tell us how it does what it does. Ironically, we may succeed in building machines that mimic human abilities yet still not understand those abilities. Much of contemporary work in AI uses hybrids of the GOFAI and Connectionist approaches. ■

Suggested Reading

Valentino Braintenberg, *Vehicles: Experiments in Synthetic Psychology*.

Daniel Crevier, *AI: The Tumultuous History of the Search for Artificial Intelligence*, chapters 1–2.

Home Page of The Loebner Prize in Artificial Intelligence, http://www.loebner.net/Prizef/loebner-prize.html.

Questions to Consider

1. You are playing Turing's imitation game, trying to ask questions that will reveal the gender of a man and a woman on the other side of a door. Each is trying to convince you that "she" is the woman. What questions would you ask in the attempt to reveal who was who?

2. Someone is trying to determine whether the communications received from you (as text on a screen) are communications from a well-programmed machine or from a human being. How would you convince this person that you are a human being?

Brains and Computers

Lecture 16

Is the brain like a computer? In this lecture, I'll examine brains and computers side by side, at different scales, from their smallest components to their overall architectures. The final conclusion is that brains are not very much like computers, but the examination that leads to that conclusion has a lot to tell us about both.

At the ground level, computers work in terms of binary digits and logic gates. Brains are built of neurons, very different and much more complex structures. Starting with a rough functional outline, a picture of the complexity of neurons is drawn by adding crucial details. Contemporary computers use a *von Neumann architecture* of central processor and memory. A quick tour of the brain shows its history and its structure and highlights important and functionally distinct areas. At this level, too, the brain is very different and significantly more complex than any computer. Do neural nets offer computers that are closer to brains? The answer is yes, but neural nets are still not very close. The lecture closes with a final contrast between how much we know about computers and how little we know about the brain.

Computers are functionally defined and, in principle, could be either analog or digital. An example of the contrast is that between old-style vinyl records (analog) and contemporary CDs (digital). Formally, a digital computer works in terms of numbers that can be expressed by terminating decimals: .134, for example. Analog computers use the full continuum of real values between 0 and 1, which includes those numbers that cannot be expressed as terminating decimals. Analog computers were constructed early on, but digital computers turned out to be easier and cheaper to construct from available components.

At the ground level, digital computers shuffle binary digits through gates. The fundamental chunk of processing is a simple on-off switch: 1 or 0. Three switches together give us eight possibilities: a byte of information.

The fundamental form of basic information is binary coding. The fundamental form of processing is the logic gate. For values of P and Q, truth tables give the value of connectives:

P	Q	P AND Q
T	T	T
T	F	F
F	T	F
F	F	F

P	Q	P OR Q
T	T	T
T	F	T
F	T	T
F	F	F

With T and F changed to 1 and 0, we get the Boolean connectives:

P	Q	P AND Q
1	1	1
1	0	0
0	1	0
0	0	0

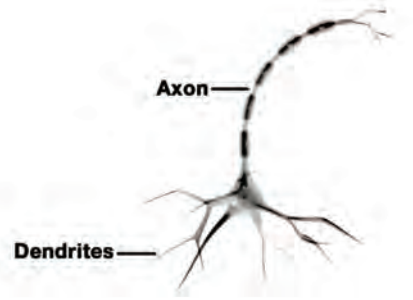
P	Q	P OR Q
1	1	1
1	0	1
0	1	1
0	0	0

At about the time of Russell and Whitehead, it was shown that the Boolean connectives can be constructed in terms of just one: the NAND gate ("not both"). At the fundamental level, all digital computing is done in terms of 1s, 0s, and NAND gates:

P	Q	P NAND Q
1	1	0
1	0	1
0	1	1
0	0	1

Virtually all programming in digital computers is serial programming, applying just one rule at each step, one step at a time. Multitasking is an illusion created by the phenomenal speed at which contemporary computers step through their paces.

At the ground level, brains work very differently. Let's start with a simple picture and make it progressively more complex. The neurons of the brain look something like a tree.



The roots are *dendrites*, which take in information from other neurons. The trunk of the neuron, which transfers the information, is the *axon*. The branches pass information on to other neurons through *synapses* at the tips. In the simplest picture, a neuron fires when its inputs reach a certain threshold. The neuron is an analog-in/digital-out device.

The discovery that neurons function electrically was a breakthrough in the 1800s, but the process is actually electrochemical rather than purely electrical. One clue is the fact that the electrochemical process operates much more slowly than a purely electrical system would. In the process of transferring an electrical potential down the axon, potassium, sodium, and calcium ions leave and enter the neuron. Neurons do not simply fire or not fire as electrical switches do. What changes is the rate of firing. There are many different kinds of neurons, 5 in the retina alone and somewhere between 50 and 500 in the brain as a whole. Information transfer at the synapses is chemical, involving at least 20 major neurotransmitters (dopamine, serotonin, and acetylcholine, for example).

The discovery that neurons function electrically was a breakthrough in the 1800s, but the process is actually electrochemical rather than purely electrical.

Much of what we know about neurons comes from the study of just one: the *giant axon* of the squid. Squids are easy and cheap to work with but have only 300 neurons. Your neocortex contains tens of billions of neurons, comparable to the number of stars in our galaxy.

At the highest level, contemporary computers use the *von Neumann architecture*. What distinguishes the von Neumann architecture is the realization that the memory unit can do two different jobs. Because of the flexibility of binary coding, a number may encode either a bit of data or an instruction. The computer memory can, therefore, contain both data and a program for operating on the data.

Corel Stock Photo Library.

Squid neurons are so large, you can see them with the naked eye.

The basic structure of every modern computer consists of:

- A memory unit for storing data and programming.
- A control unit for retrieving data and performing operations on data.
- An arithmetical side unit to help.
- Devices to facilitate input and output.

In serial programming, computer operation amounts to "fetch, operate, and store."

At the highest level, brain architecture is a matter of specialized functions in specialized areas. The brain is a massively parallel processor rather than a serial processor. The modularity of the brain is a remnant of its evolutionary history. At the core is the *reptilian brain*, which includes smell, movement, and the first reception stations for visual stimuli. The *mammalian brain* is built on top of the reptilian and includes the limbic system. It is the cortex that is most developed in humans, a wrinkled exterior that maximizes surface area.

Brain anatomy is divided into two hemispheres. The left hemisphere controls input and output for the right side of the body and the right side of the visual

field. The right hemisphere controls input and output for the left. Each hemisphere is divided into occipital, parietal, temporal, and frontal lobes. Language abilities are generally on the left, musical appreciation on the right. Broca's area controls production of speech. Wernicke's area controls comprehension of speech.

Our knowledge of brain functioning at the highest level comes from lesion studies, such as those recorded for Phineas Gage, and contemporary scanning technologies, including CT scans, PET scans, and MRIs. Several qualifications must be added to this quick tour of brain anatomy. Different brains function differently. For example, language is processed in the left hemisphere in 94 percent of right-handed people but only 73 percent of left-handed people. As indicated in earlier lectures, the brain can be amazingly plastic, recruiting different areas for varied purposes.

Are Connectionist models of neural nets closer to real brains? Artificial neural nets are more explicitly based on the analogy of real brain functioning, and they can approximate learning and pattern recognition better than traditional GOFAI systems. Nonetheless, neural nets are still biologically unreal. Artificial neural nets are trained using backpropagation of errors. We have no evidence that real neurons work in that way. Real neurons come in a wide variety. Neural nets come in just one flavor. Real neurons work in terms of changing rates of firing. Artificial neural nets do not. Real neurons actually grow, making new connections. There is no analogue in neural nets to the "chemical soup" of the brain. Although neural nets are theoretically analog devices that work in parallel, almost all contemporary work in neural nets is done using serial programming on digital computers.

We can point to one additional major difference between computers and brains. Because we design them, we understand how computers work at every level. We are ignorant of brain function at many levels. We have some grasp of brain function at the level of the neuron. We have a grasp of the brain at the largest scale, derived from lesion and scan studies. We are almost entirely ignorant of how the brain works at the crucial middle level: the level of networks of connected neurons. ■

Suggested Reading

Francis Crick, *The Astonishing Hypothesis: The Scientific Search for the Soul*, chapters 7–8.

Questions to Consider

1. Consider the following thought experiment: What we know about the human brain we know (a) at the lowest level—the level of the single neuron—and (b) at the highest level—the overall organization of major areas in the brain. We are largely ignorant of how organized complexes of neurons operate between those levels.

2. Suppose computers were something we discovered rather than built. Suppose that what we knew about these discovered computers was at similar levels—at (a) the lowest level of the NAND gate and (b) the highest level of how pieces were organized inside the case.

3. From those elements, would we be able to figure out how computers worked?

4. What would our strategy be for decoding computer function?

5. Do you think figuring out how computers function would be easier or harder than figuring out how the human brain functions? Why?

Attacks on Artificial Intelligence

Lecture 17

In this lecture, I want to concentrate on two primary critics of artificial intelligence: philosophers Hubert Dreyfus and John Searle. Their philosophical backgrounds are very different. Their critical arguments are different, but their target is the same.

Previous lectures have outlined the history of work in artificial intelligence (AI) and have emphasized some of its promise. But AI also has serious critics. Some critics target particular strategies or assumptions that have been current in the AI project. Others offer deeper arguments intended to challenge the entire idea of AI.

Dreyfus and Searle have been two of the strongest philosophical critics of AI. Dreyfus emphasizes that human intelligence is embodied intelligence. It involves recognition of relevance, the ability to shift attention, the disentangling of contextual ambiguities, and the use of rough-and-ready categories. Each of these abilities appears on Dreyfus's list of things "computers can't do," but his critiques are perhaps best construed as criticisms of a particular line of AI research at a particular historical period.

Searle's objections are stronger, offering a critique of artificial intelligence in general and in principle. Searle's core argument is the thought experiment of the Chinese room, described in detail in this lecture. How can genuine understanding consist of pure symbol-pushing? How can one get real meaning from mere mechanics? This lecture considers a range of responses to Searle from proponents of AI.

Hubert Dreyfus offers a list of things "computers can't do." Dreyfus first attacked the idea of artificial intelligence in the abusive report "Alchemy and AI," developed under the auspices of the RAND Corporation. Dreyfus's attack was further developed in later books: *What Computers Can't Do* and *What Computers Still Can't Do*. At the core of his attack is a list of four challenges:

- Humans show "fringe consciousness": a latent awareness of things in the background to which we can shift attention when needed. That shift is something computers can't do.

- We can distinguish between relevant and irrelevant factors. For the computer, everything is of the same importance.

- We can ignore contextually irrelevant senses of ambiguous words, such as *bank*. Computers cannot.

- We can recognize categories of things without lists of features, another ability that computers don't have.

How effective were Dreyfus's attacks on AI? The *frame problem* is at the core of several of Dreyfus's points but was already recognized as an important issue at the time. The frame problem can be stated as follows: Given new information, how much of the rest of my information should I update? If I learn that Bill has painted his house, I should update my information on the color of his house but not on the age of his house. One attempt to solve the problem is by building a massive database of background common sense knowledge. CYC is a 20-year, $25 million project to build such a database. Most researchers are skeptical. Another attempt is the use of *defeasible reasoning*, which works in terms of default assumptions. Defeasible reasoning avoids a number of Dreyfus's claims against programs that use strict rules.

Dreyfus's list also emphasizes an appreciation of relevance and the use of rough-and-ready categories. The development of categories is precisely what neural nets are good at. Neural nets can also be interpreted as answering central challenges of relevance.

Dreyfus's rhetoric continues to suggest that the problems he lists are problems in principle that will inevitably face any program of AI. Early claims for AI were over-hyped, but progress in AI and Connectionism make it clear that Dreyfus's attacks were over-hyped as well.

John Searle's attack on AI is deeper and more fundamental. *Weak AI* is the idea that computers can be used as a tool for understanding the mind. *Strong AI* is the claim that a properly programmed computer would *be* a mind. Searle's target is strong AI. Roger Schank designed a program to answer inexplicit questions about stories. Enthusiasts characterized a computer running the program as "understanding" the story.

Searle's counterargument is the Chinese room thought experiment. A story in Chinese is fed into a slot in the room. Questions in Chinese are fed in later. Searle envisages himself as a man in the room who consults an enormous book of rules for manipulating the symbols that are fed in the slot, eventually constructing other sets of symbols that he slips through the output slot. The man is doing exactly what a computer could do. Does he understand Chinese? No, says Searle, and the computer doesn't either.

Searle's attack is deeper than Dreyfus's. The argument can be generalized as an argument against machine instantiation of any cognitive state at any time. Any computer is a syntactic engine, operating on the form of symbols alone. The Chinese room seems to show that syntax is insufficient to give us meaning, or semantics. Some years later, Searle offered a further reading of the argument. It is we who read things as input and output, and it is we who add meaning to what the machine does.

It is we who read things as input and output, and it is we who add meaning to what the machine does.

What do proponents of AI say in their defense? Some say that the man in the room does not understand Chinese, but the system as a whole does. Searle asks: What does a physical room add? Some say the Chinese room shows the difficulties of disembodied intelligence. Searle's response is that the problem remains even if the room is inside the head of a giant robot and some of the symbols are perception reports or motor commands.

Some have emphasized the complexity of what is being envisaged and how impoverished the room analogy is. Paul and Patricia Smith Churchland say that Searle's Chinese room uses a simple picture to pluck at our intuitions

as to what would count as understanding. The Churchlands offer an analogy regarding light. A skeptic of the claim that light is electromagnetic radiation waves a magnet in a darkened room and says, "See? No light."

Searle's argument does indicate a deep conceptual problem. The problem is how meaning or semantics ever arises from something that is not itself intrinsically semantic or meaningful. Unlike Dreyfus's attack, Searle's would apply to any kind of machine. Neural nets offer no protection from the Chinese room. Philosopher Ned Block offers an extension of Searle's argument: the Chinese nation, in which people with flags perform precisely as the neurons of your brain do. Would the crowd of people that respond as your brain responds have the cognitive properties of your brain? Would the crowd understand English, even if no one in the crowd did? We do not know *how* semantics can emerge from something smaller and non-semantic, but we know that it *can*. Your brain is living proof. If the brain can do it, why not a machine? Searle's argument shows us some deep conceptual mysteries but does not thereby prove artificial intelligence to be impossible. ■

Suggested Reading

Patricia Smith Churchland and Paul Churchland, "Could a Machine Think?" *Scientific American* 262 (companion piece to Searle's "Is the Brain's Mind a Computer Program?").

John Searle, "Is the Brain's Mind a Computer Program?" *Scientific American,* 262 (companion piece to the Churchlands' "Could a Machine Think?").

———, "Minds, Brains, and Programs," *Behavioral and Brain Sciences* 3.

Questions to Consider

1. Reflect further on the Chinese room thought experiment:

> John Searle claims that the man in the Chinese room is merely shuffling symbols according to rules. The man doesn't understand Chinese.

Some of Searle's respondents have claimed that it is not the man inside the room but the system as a whole that understands Chinese.

What do you think of that response to Searle?

2. Reflect further on the Chinese nation thought experiment:

Ned Block envisages billions of people exchanging messages in precisely the way your neurons do. You understand English. Would that system of people understand English? If not, what is it that your brain has that a system of people does not?

Do We Have Free Will?

Lecture 18

Throughout the course, I've been talking about minds and machines. But there has been a philosophical issue brewing in the background of that discussion, an issue I want to take on directly: the issue of free will and Determinism.

Every event in the universe is the effect of previous causes in accordance with physical law. Every action I take is an event in the physical universe. Therefore, each of my actions must be dictated by previous causes, including causes before my birth. If so, how can my actions be genuinely free? How can I be held morally responsible for things over which I evidently have no control? This is the core of the Determinist argument against free will, an argument that has divided philosophers since the Stoics. Despite claims to the contrary, quantum mechanics offers no convenient loophole for free will.

Once we understand what free will really is, the Compatibilist says, we will see that free will and Determinism are compatible after all.

This lecture outlines a variety of Compatibilist approaches to the problem. Once we understand what free will really is, the Compatibilist says, we will see that free will and Determinism are compatible after all. On a classic Compatibilist analysis, freedom is simply the ability to do what you want, free from coercion. If so, you may act freely even if your actions are the result of causal laws and earlier events. But that classic account does not prove fully adequate to our concept of freedom. This lecture examines Harry Frankfurt's account of freedom in terms of second-order desires and concludes with a response to the Determinist argument in terms of free will in context.

The problem of free will and Determinism is an issue that just about everyone has raised in his or her own thinking. The core of the philosophical problem is a clash between two basic concepts. On one side are our concepts of ourselves as facing options and taking choices. Our concepts of moral

responsibility depend on these actions, as well. On the other side of the clash are our concepts of a universe that operates in terms of natural laws. Why does a particular event occur? It occurs because of earlier events in accordance with natural law.

In 1924, Clarence Darrow used the problem of free will and Determinism in defending Nathan Leopold and Richard Loeb. Darrow was successful; the two men were not hanged but were sentenced to life imprisonment. The problem of free will and Determinism has a long history of philosophical disagreement.

Library of Congress, Prints and Photographs Division.

Darrow's successful defense using the problem of free will spared two men's lives.

- The Stoics thought that Determinism had to be true and that effective free will was, therefore, an illusion.
- The Epicureans thought free will had to be true, and therefore, Determinism had to be false.
- In Kant's *Critique of Pure Reason*, the issue appears as one of the *antinomies*—a conceptual conflict that is rationally irresolvable.
- William James comes down on the side of freedom but renounces any attempt to prove free will.
- Jean-Paul Sartre builds his Existentialism around a commitment to freedom but gives no explanation of how it is possible.

The core of the issue is the Determinist's argument that free will cannot exist. A traditional form of the Determinist argument goes as follows:

- Everything in the universe happens because of earlier events in accordance with causal law.
- My choices and decisions are events in the universe.

- They, therefore, happen as they do because of earlier events—events even before my birth—in accordance with causal law.

- How, then, can I be said to have made a free choice? How can I be said to have acted freely? How can I be held ethically responsible for my actions?

Although the argument is usually concluded with this flourish of rhetorical questions, we bring the argument into the light of day if we replace the last step with the claim it really is: I, therefore, have no free choice. I cannot act freely and cannot be held ethically responsible for my actions.

The whole problem of free will and Determinism can now be encapsulated in the question: What should we think of that argument? Is the argument valid—does the logic go through? The Stoics thought it was valid, that the premise was true, and thus, concluded that we have no free will. The Epicureans thought the argument was valid, but the conclusion was false; therefore, the premise must be false, as well—the universe does *not* operate entirely in terms of Deterministic natural law.

Contemporary physics agrees with the Epicureans. According to standard interpretations of quantum mechanics, the fundamental laws of physics are statistical rather than Deterministic. Quantum mechanics is perhaps the best confirmed scientific theory in history, but the theory entails the idea that some events happen with no cause, such as the decay of a particular uranium atom at a particular time. Does that leave a loophole for free will? Not really. It would merely replace one thing over which I have no control—Deterministic events in a Deterministic universe—with another—random quantum events in my brain.

The Compatibilist approach offers another way out. Perhaps freedom and Determinism are not incompatible after all. In a classic Compatibilist account, freedom is defined as having the ability and power to do what you want. The conditional account is as follows: I wanted to take the road to the left and did so. To say that I was also free to take the road to the right is to say that I would have taken that road if I had wanted to. On such an account, it could be true *both* that I freely chose a particular road *and* that my

choice was the result of previous events. Compatibilism has an explanation for where Determinism goes wrong: It confuses causality with coercion.

The conditional account does not appear to be completely adequate. What about cases in which someone could not have chosen otherwise? Addiction offers a counterexample to the conditional account. The mad neuroscientist offers another: He manipulates his victims by manipulating their desires.

New forms of Compatibilism offer new views of free will. Does free will demand alternative possibilities? The thought experiment of Harry Frankfurt suggests that it might not.

- John W. Oswald decides to shoot President Keneagan. He plots and plans, aims, and pulls the trigger.
- The neurophysiologist Professor Moriarty has implanted a device in Oswald's brain that will force him to kill Keneagan if he happens to change his mind about his desire to do so.
- Oswald does not change his mind, the device is not activated, and he does pull the trigger.
- Did he act freely in shooting Keneagan? Yes. But he could not have done otherwise.

Frankfurt uses the case of Oswald to argue for a hierarchical account of free will. Frankfurt introduces *second-order desires*, that is, desires to have certain desires. The self-reflection required for second-order desires is characteristic of humans and is central to free action. An act is free just in case it accords with our second-order desires. Frankfurt's account seems to give the right answer in the Oswald case, but it seems to fail in the case of addiction so invidious that it comes to affect even the desires an addict *wants* to have.

Attention to context offers new options for Compatibilist accounts. The Determinist argument assumes that freedom of an action is a property that an action either has or does not have, regardless of context. That assumption may be wrong. The concept of *smooth* offers a comparison case. Whether

something counts as smooth depends on context. Concepts of freedom and free choice may also depend on context.

The Determinist argument may be switching context behind a context-sensitive term. The first steps of the Determinist argument set a context in which the demands for free action are extremely high. In that extreme sense, perhaps we are not free. That doesn't mean we might not be free in a much more normal sense, but the Determinist's conclusion suggests we are not—a conclusion that I have no choice as to what I'll eat for dinner or who to vote for. By switching context halfway through the argument, the Determinist has essentially changed the meaning of the term and, thus, introduced the logical fallacy of ambiguity. The Determinist has proved something about freedom in an extreme and unrealistic sense but has fallaciously concluded something about freedom in the very different sense that we really care about. What we really need to understand is the central concept of freedom. ■

Suggested Reading

Daniel Dennett, *Elbow Room: The Varieties of Free Will Worth Wanting*, chapter 1.

Patrick Grim, "Free Will in Context: A Contemporary Philosophical Perspective," *Behavioral Sciences and the Law* 25.

Questions to Consider

1. Clarence Darrow used the problem of free will and Determinism in 1924 to argue that Nathan Leopold and Richard Loeb should not be given the death penalty.

2. "Is Dicky Loeb to blame because of the infinite forces that conspired to form him, the infinite forces that were at work producing him ages before he was born … ?"

3. What if Darrow had used the same defense in arguing that Leopold and Loeb should be found not guilty and go free? Would the argument have been weaker in that case?

4. The philosophical issue of free will is often expressed as if one either has free will or does not. But we also speak of the possibility of being more or less free. Do you think that some of the people you know are less free than others? In what ways? For what reasons?

Seeing and Believing

Lecture 19

At this point in the lecture series, I want to train all our resources—the resources of philosophy, the resources of psychology, and of the neurosciences—on the central issue of consciousness. In this and the next lecture, I'll be talking about some of the things we know about consciousness, about perceptual consciousness, in particular.

Do we see only what we want to see? This lecture uses a range of data from psychology to explore the ways in which our conscious experience is shaped by background beliefs and expectations. The issue also has important implications for our system of justice. How reliable is eyewitness testimony?

This lecture also links these psychological data to a wider philosophical issue of theory-laden perception. Because our beliefs can influence what we see, some thinkers have argued that rational scientific change is impossible and that scientific objectivity is a myth.

Although following that line of argument can reveal some important ideas, the lecture argues that extreme conclusions are ultimately unjustified. The influence of belief on perception, though pervasive, has demonstrable limits. Both our individual minds and our collective science manage information by carefully balancing the influence of immediate perception and background information.

The psychological evidence is ambiguous regarding the impact of desires on perception. But it does support something close: that you tend to see what you expect to see. This lecture focuses on how background beliefs can shape perception. Eyewitness testimony will serve as a case in point. This lecture also explores the philosophical implications of theory-laden perception for the idea of objectivity in science.

A number of classical studies show that expectation influences what we see. Jerome Bruner showed that altered cards (a red ace of spades, for example,

or a black jack of diamonds) are more difficult to process. In another experiment, subjects are told to count the number of times that people in a video pass a basketball. About half of the subjects fail to see the person who walks through in a gorilla suit. The *cocktail party effect* is an auditory example of the force of attention. You filter out the conversations around you … until you hear someone mention your name.

The tendency of a first interpretation to last, resistant to change, is called *perceptual persistence*. Subjects who are shown the following images in left to right order starting from the top tend to see a man's face well into the images in the second row. Subjects who are shown the images in the reverse order tend to see a woman's figure well into the images of the first row

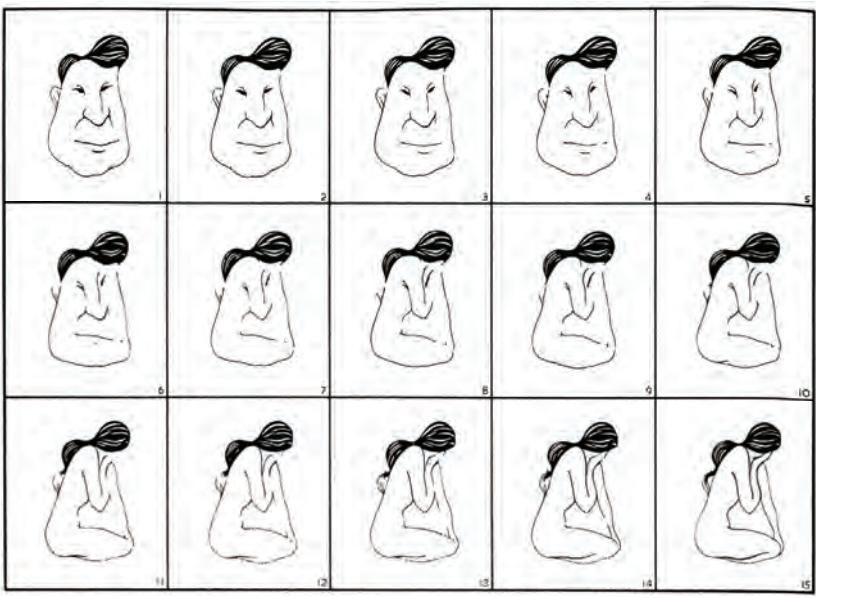


Figure entitled "The 'Man' and 'Girl' set of ambiguous figures" from the article published in *Perception & Psychophysics*, Volume 2, Issue 9 (1967) entitled "Preparation of ambiguous stimulus materials" by Gerald H. Fisher.

Change blindness labels the fact that we fail to see changes in a scene. This phenomenon is demonstrated by a range of experiments regarding continuity in both video clips and still images.

These results show that perception is an active *pursuit* of data rather than a passive reception of data. The active pursuit of perception can be expected to access a variety of cognitive resources, including background beliefs, expectations, and interpretations. These results accord perfectly with the evolutionary perspective on perception outlined in Lecture Eleven.

The fact that perception accesses background cognition can lead to errors. A particularly worrisome case of this is in eyewitness testimony. The popular conception is that eyewitness testimony is perhaps necessary and almost certainly sufficient for proof of guilt. Juries in British cases rendered a verdict of guilty in 74 percent of cases in which the only evidence was a single eyewitness.

The fact that perception accesses background cognition can lead to errors. A particularly worrisome case of this is in eyewitness testimony.

How good is eyewitness testimony? This answer is: not particularly good. In classroom experiments, I have found that people do well in reporting the number of assailants who enter the room and what they say. They do worse in identifying who said what and in identifying the perpetrators from a photo lineup. *Twelve Angry Men* is a wonderful dramatization of a jury deliberation that turns on some of the many ways that eyewitness testimony can go wrong. Will a real crime be reported more accurately? Studies indicate that witnesses give *less* accurate reports of violent events.

One thing witnesses do tend to report accurately is the type of weapon used. This phenomenon is called *weapon focus*. In a study by C. Johnson and B. Scott, volunteers were told to wait in a reception room. Some volunteers hear a conversation in a nearby room. A man then comes out of the room with a pen in his hands, which are covered with grease. Other volunteers hear a loud argument in the other room. A man then comes out with a letter opener in his hands, which are covered with blood. In the case of the pen and

the grease, 49 percent of subjects were able to identify the man later from a photograph. In the case of the letter opener and the blood, only 33 percent could later identify him.

How initial questions are phrased can influence later testimony. In a classic study by Elizabeth Loftus, a leading expert in the subject, people are shown a film of a traffic accident. Half are asked, "About how fast were the cars going when they hit each other?" The other half are asked, "About how fast were the cars going when they smashed into each other?" A week later, twice as many subjects in the second group, the "smashed" group, reported seeing broken glass as in the first group, the "hit" group.

Library of Congress, Prints and Photographs Division, LC-USZ62-1433.

Benjamin Franklin's experiment with electricity took place in 1752.

The fact that perception can be cognitively loaded has been influential in 20th-century philosophy of science. If we take this statement far enough, it appears to threaten the very idea of scientific objectivity. Thomas S. Kuhn effectively challenged the view that science is cumulative. When examined in historical detail, science proceeds by revolutions in which new theories *replace* old theories. According to Kuhn, science at any point is ruled by a particular paradigm. The paradigm, in turn, colors our interpretation of experiments. For Kuhn, perception is theory-laden.

If interpretation of experiments is theory-laden, people with different paradigms will simply read the experiments as supporting their own theories. N. R. Hanson poses the problem of objectivity. Let us consider Johannes Kepler. Imagine him on a hill watching the dawn. With him is Tycho Brahe. Kepler regarded the sun as fixed: it was the earth that moved. But Tycho followed Ptolemy and Aristotle in this much at least: the earth was fixed and

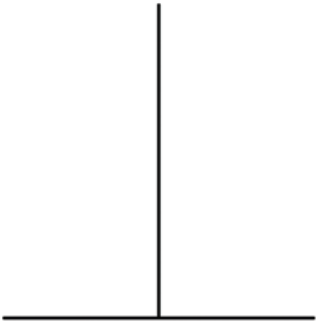
all other celestial bodies moved around it. *Do Kepler and Tycho see the same thing in the east at dawn?*" Sometimes Kuhn and his followers suggest that there can be no rational transition from one paradigm to another. The shift is more like a religious conversion.

The Teaching Company Collection.

Isaac Newton.

The strong Kuhnian argument rests on two premises. The first premise is that our beliefs form a seamless, unified whole guided by a single paradigm. This is Kuhn's *holism*. The second premise is that all perception is thoroughly theory-laden. If one accepts both premises, it will follow that no experiment will ever settle any issue. Science will be hopeless. The conclusion of that argument is too strong. Were it true, the crucial experiments of Galileo, Franklin, and Newton and the tests of Einstein's theory of relativity would have been futile because no experiment could ever be convincing.

Both premises of the strong Kuhnian argument are open to challenge. Scratch the surface of any scientific discipline and you will find areas of well-cemented opinion, areas of open question, and wide areas of disagreement. The idea that background belief and expectation bias perception entirely and always can be shown to be false. Consider the *top hat illusion*, in which the vertical line appears longer than the horizontal:



If you measure the two lines, you will see that they are exactly the same length. You can construct your own version of the illusion by using a ruler to draw the lines. Although you know the lines are the same length, the vertical line will still *look* longer. If all belief influenced all perception, that would not be true. Although theory may influence perception, it is not true that all perception is thoroughly theory-laden.

The lessons of this lecture apply to two systems that are both knowledge-guided and knowledge-gaining. One is local and individual. The system is your brain: an individual system that is both knowledge-gaining and knowledge-guided. The other system is culture-wide: The sciences are a system that is both knowledge-guided and knowledge-gaining.

This mutually reinforcing dynamic in which perception influences belief and belief influences perception has its dangers. If belief were swamped by immediate perception, we would lose the benefits of background knowledge. If perception were swamped by belief, we would lose the flexibility needed to deal with change. Both systems can function only if background belief and perception remain in equilibrium. ■

Suggested Reading

Jerry Fodor, "Observation Reconsidered," *Philosophy of Science* 51.

N. R. Hanson, "Observation," *Patterns of Discovery*.

Elizabeth F. Loftus, *Eyewitness Testimony*, chapter 4.

Reginald Rose and Sidney Lumet, *Twelve Angry Men* (video).

Visual Cognition Lab, University of Illinois, http://viscog.beckman.uiuc.edu/djs_lab/demos.html (shows the unseen gorilla video mentioned in the lecture, together with wonderful exercises regarding change blindness).

Questions to Consider

1. N. R. Hanson writes:

> Let us consider Johannes Kepler: Imagine him on a hill watching the dawn. With him is Tycho Brahe. Kepler regarded the sun as fixed: it was the earth that moved. But Tycho followed Ptolemy and Aristotle in this much at least: The earth was fixed and all other celestial bodies moved around it. *Do Kepler and Tycho see the same thing in the east at dawn?* In what ways might the answer to Hanson's question be "No, they do not see the same thing"? In what ways might the answer be "Yes, they do see the same thing"?

2. Have you experienced cases in which your desires or expectations shaped your perception? In what way was your perception influenced? How did you come to see the influence of desire or expectation?

Mysteries of Color

Lecture 20

In this lecture, I want to examine our consciousness of color from several different directions. There are things we know from physics, psychology, and the neurosciences regarding color and color perception, but there are also mysteries that remain.

Is color an objective feature of the world or something that exists only in subjective experience? In his famous experiment, Newton broke white light into its component colors through a prism, laying the groundwork for our understanding of color in terms of wavelengths. Goethe wrote a direct response, claiming that Newton was wrong about color. According to Goethe, color exists only in the mind. Who is right about color?

In this lecture, data from both psychology and the neurosciences are used to reveal the complexities of color perception, but both lead to a basic philosophical question regarding subjective experience: Do you see the same color I do? Could it be that you and I have learned to apply color terms in similar ways but nonetheless have radically different subjective experiences when we see something red? We will explore this philosophical question using Locke's thought experiment of the inverted spectrum on one side and Daniel Dennett's Chase and Sanborn thought experiment on the other. Work by neurophysiologist Stephen Palmer can help us close in on the problem … but not quite solve it.

Physics, psychology, and the neurosciences give us some knowledge about color and color perception. Nonetheless, mysteries remain. Primary among these are philosophical questions regarding color qualia: How do I know that your subjective experiences of red things are the same as mine?

Two radically different views of color are represented in the work of Newton and Goethe. Newton's experiment of 1666 lays the groundwork for our contemporary understanding of color in terms of wavelength. Goethe claims that Newton's approach is inadequate. Colors exist only in the mind.

Philosophical controversy concerning the nature of color continues.

- On an objectivist view, colors are properties in the world.
- On the subjectivist view of Galileo, Descartes, and Locke, colors are mind-dependent secondary qualities.
- On a relational approach, colors exist only in relation to a color perceiver.
- On a dispositional approach, colors are tendencies to produce subjective experiences in an observer.
- What is it about an object that produces the subjective experience? Perhaps it is something about the microstructure of the object. Here, we are back to objectivism again.

Everyone agrees that color consciousness and color perception involve both objective features of the world and the experience of observers.

Our investigation can start with the psychology and neuroscience of color perception. Two central facts show that color perception is more than wavelengths.

- *Color constancy* is the fact that colors are perceived to be the same despite changes in wavelength.
- *Color contrast* is the fact that colors depend on contrast in context. A startling example can be found online at the site listed below in Further Resources.

Both color constancy and color contrast fit in perfectly with an evolutionary theory of perception.

Much of what we know of color vision has come from the interaction of science and technology. James Clerk Maxwell laid the foundations for color photography in 1861. His experiment showed that our visual system would

need only three types of color receptors to capture the full spectrum of colors we see. Edwin Land offered an even more startling experiment regarding color, concentrating on the contrast effect.

Our contemporary understanding of color processing is precisely in line with Land's *retinex* theory of color context. There are three types of cones in the retina, each of which uses a different pigment to respond to a different wavelength of light: long, middle, and short. Ganglion cells respond to differential input from the three kinds of cones in their receptive areas. Are short waves predominant or middle and long? Are long wavelengths more dominant than medium or short?

A basic philosophical problem arises concerning our subjective experience at the end of the process. All of our data on color perception are behavioral. But couldn't your behavioral response be identical to mine yet our subjective experiences be radically different? John Locke phrases the problem in terms of the inverted spectrum. Might not the spectrum of your color experience be directly inverted from mine? The technical term for subjective qualitative experiences is *qualia*. If qualia exist in this sense, they pose a real problem for both Behaviorism and Functionalism.

Remember your first taste of beer? What awful stuff! What has changed, the taste of beer or how you react to the taste? Is there really any difference between those explanations?

Daniel Dennett has made one of the most valiant attempts to deflate the notion of qualia. One of his thought experiments involves Chase and Sanborn. Mr. Chase says, "The qualia are the same as they always were, but my tastes have changed." Mr. Sanborn says, "The qualia have changed … the coffee tastes different to me now." Dennett asks whether any difference really exists here. Remember your first taste of beer? What awful stuff! What has changed, the taste of beer or how you react to the taste? Is there really any difference between those explanations?

The psychologist Stephen Palmer suggests that behavioral tests can reveal an inverted spectrum. Color can be conceived of in terms of three dimensions

rather than just one: hue, brightness, and saturation. Relationships between color hues are represented in the color circle. The color wheel adds the parameter of saturation. The color sphere adds intensity. Yellow is brighter than blue even when they have the same intensity. If the spectrum you see were inverted from mine, you would think that a certain blue was brighter than a certain yellow. I would not. In order to get the problem going, we would need to invert all the qualities of the color sphere.

Palmer also proposes that in some cases, the problem may not be merely a thought experiment.

- *Protanopes* have a gene that gives their long-wavelength cones the medium-wavelength pigment by mistake.

- *Deuteranopes* have a gene that gives their medium-wavelength cones the long-wavelength pigment by mistake.

Both occur in the population. We should, therefore, expect some proportion of people to have both problems. *Reverse trichromats* would have pigments reversed between these two types of cones. For such individuals, Palmer proposes, behavioral tests for an inverted spectrum would be impossible.

The big philosophical problem at issue can be posed in terms of four examples. Mr. Chase and Mr. Sanborn characterize what has happened to them in different ways. But is there any real difference there? Does beer taste different now, or have you just come to like the taste? Is there any real difference between those alternatives? In a third stage of wearing inverted lenses, mentioned in Lecture Twelve, "everything seems right again." Is that because sensory experience has flipped back or because you have learned to navigate effectively? Is there any difference between those alternatives?

Suppose the pigments in your long-wavelength cones and medium-wavelength cones had been reversed from birth. Would that make a real qualitative difference in your world? If your tendency is to say no to these questions, you are leaning toward something like Functionalism. If your tendency is to say yes, you are leaning in the other direction. You take qualia seriously. ■

Suggested Reading

Donald D. Hoffman, *Visual Intelligence: How We Create What We See*, chapter 5.

Oliver Sacks, *An Anthropologist on Mars: Seven Paradoxical Tales*, chapter 1.

Duke University, Neurobiology, Laboratory of Dale Purves, M.D., *Color Contrast: Cube*, http://www.neuro.duke.edu/faculty/purves/gallery9.html (shows the color contrast experiment mentioned in the lecture; the experiment appears with some additional discussion at http://discovermagazine.com/2004/feb/neuroquest).

e-Chalk Optical Illusions, http://www.echalk.co.uk/amusements/OpticalIllusions/illusions.htm (interactive color illusions).

http://www.lottolab.org/.

Questions to Consider

1. Newton says that white light is composed of a combination of colors, which are frequencies of light. Goethe says that color exists only in the mind. In that dispute, do you tend to agree with Newton or Goethe?

2. You wake up one morning and things seem to have changed color: Things you thought of as red now look green. How would you tell whether:

- How you currently perceive things has changed, or
- Your memory of how things looked before has changed?

3. Daniel Dennett seems to suggest that there is no real difference between the two alternatives. Do you agree with Dennett or not?

The Hard Problem of Consciousness

Lecture 21

Conscious experience is what you're having right now. But when it comes to understanding what that conscious experience really is and how it really works, nothing could be more mysterious.

If there is a defining problem in philosophy of mind today, it is the problem of accounting for our subjective experience. David Chalmers calls this the "hard problem of consciousness." Thomas Nagel presses the hard problem in asking, "What is it like to be a bat?" Frank Jackson uses the example of color-blind Mary to argue that there is something about our subjective conscious experience that cannot be caught in the net of scientific knowledge. The hard problem of consciousness challenges contemporary Functionalism with a renewed Dualist question: How could any physical or functional account explain the taste of pineapple, the felt chill of a cold winter's morning, or why things look, taste, or feel certain ways at all?

Is a Functionalist account of phenomenal consciousness possible, or is consciousness a counterexample to the Functionalist claim that all mental states are functional states? This lecture explores the hard problem from both the Functionalist and the Anti-Functionalist side. The thought experiment of zombies is used on both sides of the issue.

In a conceptual form, the thought experiment of zombies is a straightforward argument against Functionalism. In the guise of evolutionary biology, however, it raises questions that may ultimately favor a Functionalist approach.

Consciousness is used in a number of ways. The sense at issue here is what philosophers sometimes call *phenomenal consciousness*, the qualia of our subjective experience. Nothing could be more familiar than consciousness in this sense. But when it comes to understanding subjective experience, nothing could be more mysterious.

The problem of consciousness is a particularly vivid form of the mind-body problem. Let's start our exploration with a summary of some of our theoretical work to this point. According to Dualism, the mental and the physical are radically different; mental phenomena do not even occupy space. How, then, could the two interact? Epiphenomenalism, Occasionalism, and Parallelism were attempts to patch over the problem of interaction.

The alternative to Dualism is Monism. Idealism holds that there is no physical world, only a mental one. Materialism holds that everything is ultimately physical. How could mental phenomena be physical? Behaviorism offers one answer: What we are really talking about when talking about mental states is behavior. The dominant theory today is Functionalism. When we are talking about mental states, we are talking about the functional organization of organisms.

Functionalism has a number of attractive features, but a problem remains. Functionalism lays out a route for research that links a number of different fields. If mental states are functional states, we can explore their functional organization. That is the program of cognitive psychology, which has broadened into cognitive science.

If mental states are functional states instantiated in brains, we can understand them by understanding how brains work. Neuroscience is part of the research program here. If mental states are functional states, they might also be instantiated in machines. Artificial intelligence and robotics are part of the research program, as well.

Another attractive feature is that Functionalism allows us to understand the universe as a unified whole. Functionalism also allows us to see the mental as somehow different—as a matter of functional organization rather than material composition. Despite Functionalism's appeal, there may be an obstacle in the road: the obstacle of consciousness. A number of philosophers argue that Functionalism cannot account for subjective experience.

David Chalmers contrasts the "easy" problems with the "hard" problem of consciousness. The easy problems include these:

- To explain our ability to discriminate, to categorize, and to react to environmental stimuli.
- To explain how a cognitive system integrates information.
- To explain how a system can have access to its own inner states.
- To explain how a system can report on its own mental states.
- To explain how a system can focus attention and deliberately control behavior.

The easy problems are easy in the sense that they are questions of function and fit a Functionalist research program perfectly. The "hard problem" is the problem of consciousness: to explain the *subjective conscious experience* of seeing a sunset. Subjective conscious experience, says Chalmers, is something that Functionalism will inevitably leave out.

A supporting argument comes from Thomas Nagel. Nagel asks, "What is it like to be a bat?" Bats locate prey by echolocation. How does echolocation *feel*? The question is not what it is like for me to imagine myself flying at night and making high-pitched sounds. The question is what it is like for the bat to be a bat. To be a being with subjective consciousness means that there is something it is like to be that being. No matter how complete our physical and functional knowledge regarding a bat, its subjective experience will evade us. This is the mind-body problem again.

Corel Stock Photo Library.

We can see a bat, hear a bat, touch a bat, maybe even smell a bat, but we cannot know what it actually feels like to be a bat.

Another supporting argument comes from Frank Jackson. Mary is a neuroscientist. She lives 50 years in the future and knows all that a fully

developed neuroscience could tell us about color perception. Mary, however, has never seen anything but black and white. Suppose we open the door and let Mary come out into the real world. She now sees what color looks like for the first time. Jackson asks, "Has Mary learned something new?" The answer seems to be yes. Jackson argues that no understanding of color from physics or neuroscience can give us the whole story about color. Our science leaves out the subjective experience of color.

These arguments carry a punch, but what exactly do they tell us about consciousness? Chalmers, Nagel, and Jackson think these arguments support the position that consciousness is special, impossible to catch in our scientific nets. This view is clearly a contemporary form of Dualism.

If consciousness is beyond the reach of contemporary science, how are we to understand it? One answer, explored in Lecture Twenty-Two, is that we will have to stretch contemporary science to include it. Another answer, explored in Lecture Twenty-Three, is that we will *not* be able to understand it.

Is there an alternative? What do Reductive Materialists and Functionalists say in response to the "hard problem" arguments? One response is to "hope for a miracle." Perhaps we will find a physical structure or process that suddenly makes consciousness understandable. The other reply is a Deflationist response. When we deflate our concept of consciousness, it will become clear that a Functionalist account will be possible and that the hard problem isn't so hard after all.

The zombies of philosophical thought experiments are explicitly defined as beings with full functionality but without any subjective consciousness.

A further way of considering this difficult conceptual terrain is in terms of the philosophical thought experiment of zombies. The zombies of philosophical thought experiments are explicitly defined as beings with full functionality but without any subjective consciousness. Are zombies possible? Could such things really exist? If zombies are possible, it is possible for something to be functionally identical to you without being conscious. Consciousness must, therefore, be something more

than a matter of how something functions. The Anti-Functionalists win. The Functionalist must say that zombies are impossible; anything that functions as you do in all possible situations would have to count as conscious.

Zombies can also be used to raise a further question in terms of evolutionary biology. What is consciousness for? Evolutionary pressures operate in terms of the functioning of an organism. If a species of zombies were in competition with conscious beings like us, why should conscious beings win? If consciousness is part of the physical universe, it uses energy. Zombies would have an energy advantage in the evolutionary race. Why isn't the world populated by zombies, then, instead of us? One answer is that consciousness must have some special function that was selected for in evolution. The Functionalist is right after all.

Not everything in the evolutionary record, however, is a trait that was selected for. Evolutionary *spandrels* are features that just happened to come along for the ride. The color of your eyes may have been evolutionarily selected for, but the color of your blood appears to be a spandrel. If consciousness is a spandrel, then this thing we think of as most characteristic of our inner selves is just an evolutionary accident. ■

Suggested Reading

David Chalmers, "Facing Up to the Hard Problem of Consciousness," *Journal of Consciousness Studies* 2.

———, *Zombies on the Web*, http://consc.net/zombies.html.

Thomas Nagel, "What Is It Like to Be a Bat?" *The Philosophical Review* 83.

Daniel Stoljar and Yujin Nagasawa, *There's Something about Mary*, introduction.

Questions to Consider

Frank Jackson's thought experiment is as follows: The eminent neuroscientist Mary is the world's expert on the perception of color. She leaves her black-and-white world and sees color for the first time. "So that's what it looks like!" she says.

1. What is it that Mary has learned?

2. Is there any way of putting what she has learned into words?

3. Does this thought experiment show that some aspects of consciousness are beyond the reach of science?

The Conscious Brain—2½ Physical Theories

Lecture 22

The core of the hard problem is this: How are we to understand conscious experience? How could the three pounds of gray and wet cellular matter in a human skull produce *this*: the subjective realm of our experience of colors, of tastes, of smells, of feelings?

This lecture focuses on two contemporary attempts to explain consciousness within the confines of physics and neuroscience. After discovering the structure of DNA with James D. Watson, Francis Crick turned his attention to the mysteries of consciousness. With Christof Koch, Crick takes the binding problem to be central and proposes a neuroscientific theory of consciousness in terms of the synchronized firing of neurons in spatially separated areas of the brain.

Building on results in logic—the halting problem and Kurt Gödel's incompleteness theorem—mathematical physicist Roger Penrose argues that the brain can do things that no computer can. His search for consciousness leads him to look for a nonalgorithmic mechanism deep in the brain. Penrose's hypothesis is that quantum phenomena are crucial to consciousness, amplified within special structures in our neurons.

In a final section, the lecture also broaches David Chalmers's more radical suggestion as a "half theory." Chalmers proposes that we should revamp our entire scientific worldview in order to include consciousness as a fundamental principle all the way down to the smallest particles of the universe. But would even *that* give us any real understanding of consciousness?

Crick offers a theory of consciousness in terms of synchronized neural firing. After receiving the Nobel Prize for discovering the double-helix structure of DNA with James D. Watson, Crick set out to solve another mystery: the mystery of consciousness. The unity of consciousness is a central aspect addressed in the theory developed by Francis Crick and Christof Koch. Impressions from different senses—from hearing, sight, and touch—come together in a single consciousness. How does that happen? Both Plato and

Aristotle talked about the problem. Our contemporary understanding of the brain makes the problem even more perplexing. Visual color, visual motion, and visual shape are processed in different areas of the brain. How is that disunity resolved in the unity of consciousness? Within the neurosciences, this is known as the *binding problem*.

Crick and Koch's theory proposes that binding in the brain is something that occurs not in space but in time. As outlined in Lecture Sixteen, neurons always have a background rate of firing, which increases to fast and regular firing when they are stimulated. In Crick and Koch's theory, neurons in different areas of the brain bind by synchronizing their firing in the range of 40 Hertz (40 evenly spaced firings per second).

How good is the 40-Hertz theory? Crick and Koch sometimes hedge their bets, characterizing it as a theory of neural correlates to visual awareness. Could the theory be taken to explain subjective conscious experience? If 40-Hertz synchronization happened precisely when subjective experience was present, a good guess would be that it actually produced consciousness. But the "hard problem" is *how* any such process *could* produce subjective experience. That would still remain unexplained. The 40-Hertz theory is worth pursuing, particularly with an eye to understanding the binding of separate areas in the brain. Nevertheless, it does not offer an answer to the hard problem.

Roger Penrose offers a second physical theory of consciousness driven by one of the most startling results in 20th-century logic. One form of the result is Alan M. Turing's *halting problem*. In the 1930s, Turing offered a precise formal model for the concept of computations or algorithms: step-by-step procedures for finding something out. The abstract model is called a *Turing machine*. Turing also showed that there were things that no Turing machine could calculate. One example is the halting problem: No Turing machine could predict in all cases whether a Turing machine will go into an infinite loop.

Another form of the result is Gödel's *incompleteness theorem*. Axiomatic systems are like those in Euclid, taught in high school geometry. From basic principles, we prove further theorems. We can pinpoint clear desiderata in

building any axiomatic system, such as an axiomatic system for arithmetic. We want an axiomatic system that gives us the truth, the whole truth, and nothing but the truth. Gödel's incompleteness theorem proves that no axiomatic system of even minimal power can give us the truth, the whole truth, and nothing but the truth.

Gödel's incompleteness theorem proves that no axiomatic system of even minimal power can give us the truth, the whole truth, and nothing but the truth.

Penrose asks: Can human minds do something no algorithm can? Can human minds do something no formal system can? Turing and Gödel constructed those proofs, and we can understand them. We can see the limits of formal systems, Penrose says, and can see what lies beyond. Thus, our brains must be able to do something no algorithm and no computer can do. Somehow, our brains must function nonalgorithmically. This step alone is controversial. Gödel's and Turing's results are closely related. Do these results show that people have some power that machines do not? Gödel thought the answer was yes. Turing thought the answer was no.

How does formal undecidability lead to consciousness? The answer is: by way of quantum mechanics. Penrose thinks that our nonalgorithmic abilities demand intelligence and insight and that these are essentially conscious. He also thinks that free will demands consciousness. According to quantum mechanics, fundamental physical processes are nonalgorithmic and non-Deterministic.

Working with Stuart Hameroff, Penrose proposes that consciousness—nonalgorithmic and non-Deterministic—can be explained by quantum effects in neurons. Neurons have cytoskeletons composed of microtubules, with internal spaces on the order of a millionth of a millimeter. Those spaces might be small enough for quantum effects to appear. Penrose and Hameroff propose that the microtubules could amplify those effects, thereby producing consciousness.

How good is Penrose's theory? Some of the problems are in the technical machinery. Gödel's and Turing's results show things that no algorithm and

no formal system could compute, but do we really know that *we* could compute them? Quantum indeterminacy appears on the level of the very small but tends to cancel out by interference on larger scales. All cells have cytoskeletons and microtubules. Why are only brain cells conscious? Both consciousness and quantum mechanics are mysterious, but it doesn't follow that one mystery explains the other.

David Chalmers proposes a fairly wild direction in which to look for a physical theory that would answer the hard problem. Chalmers argues that we will have to change our scientific worldview to accommodate consciousness. Consciousness should be added as a further fundamental principle, operating throughout the universe. Chalmers proposes a form of *Panpsychism*, according to which everything—electrons included—has an aspect of consciousness. Everything—electrons included—has some form of subjective experience.

The proposal demands so much and offers so little. It requires a revamping of physics down to the smallest particle. But consciousness seems to appear only in higher organisms. The theory would see consciousness everywhere but would give us no more *understanding* of consciousness than we have now. ■

Suggested Reading

Francis Crick, *The Astonishing Hypothesis: The Scientific Search for the Soul*, introduction and chapter 17.

Roger Penrose, *Shadows of the Mind: A Search for the Missing Science of Consciousness*, chapter 7.

Questions to Consider

1. Suppose that Francis Crick's 40-Hertz theory of synchronized neuron firing is correct. What aspects of consciousness might that explain and what aspects would it not explain?

2. Suppose that Roger Penrose is right that the brain exploits quantum phenomena in the microtubules of neurons. What aspects of consciousness might that explain and what aspects would it not explain?

The HOT Theory and Antitheories

Lecture 23

In this lecture, I want to consider two further approaches to the problem. One of these is a very different kind of theory, the higher-order thought, or HOT, theory of consciousness. It's perhaps the strongest contender for a fully Functionalist theory of consciousness. The other approach is represented not by a theory of consciousness but by a handful of antitheories.

The previous lecture focused on physical theories of consciousness. The philosopher David Rosenthal has offered a very different theory of consciousness, a Functionalist theory designed to deflate the stories we tell ourselves. A conscious mental state, Rosenthal argues, is not one that has a special conscious "glow." A conscious state is merely a state we are conscious *of*, in much the same sense that we might be conscious of an object in the world.

This lecture also surveys antitheories of consciousness: arguments that a science of consciousness is simply impossible. Thomas Nagel argues that two radically different perspectives on the world exist side by side. One is essentially subjective, tied to a particular point of view. The other is an objective perspective, essential to science: the "view from nowhere." Because consciousness is essentially subjective, it is the one entity objective science will never be able to explain. Colin McGinn argues that there must be a physical truth about consciousness, but it will be conceptually impossible for us to grasp. Both Nagel's and McGinn's arguments are considered critically in this lecture. If there is always a possibility that we could be wrong, is there not always a presumption against theories that would shut down further inquiry?

The HOT theory is attributable to David Rosenthal. Deflationist strategy in tackling the hard problem is to look hard at the concept of consciousness and to deflate it. When we understand the concept of consciousness, we will see that the hard problem is not so hard after all. Rosenthal makes a number of distinctions in our concept of consciousness. Sometimes we speak of people

or animals as being conscious when they are awake and responding. This is *creature consciousness*. Sometimes we speak of a belief or an emotion as being conscious. This is *state consciousness*. When you are conscious of a car in front of you or a cup beside you, yours is *transitive consciousness*. We also speak of a state, intransitively, as a conscious state on its own. This is *intransitive consciousness*. The focus of the theory is on intransitive state consciousness.

What the HOT theory attempts to deflate is the notion that consciousness is a particular quality of some mental states. In that view, consciousness is an intrinsic character of a mental state. In the HOT view, consciousness is essentially relational instead. A state is conscious because of its relation to something else.

The core of the HOT theory is as follows: To say that a mental state is conscious is merely to say that there is a higher-order thought *about* it. A mental state is conscious when it is the transitive target of some other mental state. The theory is entirely Functionalist. Consciousness is, after all, a matter of the functional organization of mental states. There is an interesting parallel between Rosenthal's HOT theory of consciousness and Harry Frankfurt's hierarchical theory of free will, outlined in Lecture Eighteen.

How good is the HOT theory? It has both positive and negative points. On the positive side, the HOT theory is the closest we have to a successful Deflationary theory of consciousness. The relational characteristics that Rosenthal points out *do* seem to be characteristics of subjective experience.

On the negative side, the theory may be smuggling in what it is trying to explain through the back door. In order to work, the theory must exclude certain cases. If I have a higher-order thought about my anger because you tell me I am angry, my anger may or may not become conscious. In a refinement of the theory, Rosenthal says that the higher-order thought must link to its target in an *immediate* way. Does *immediate* mean that the lower state has to be directly felt or sensed … that it has to glow? If so, the theory may be smuggling in what it is supposed to explain.

Thomas Nagel offers an antitheory of consciousness. Nagel says that we have two different perspectives on the world. One perspective is essentially subjective—the perspective from my experience, my sensations, and my point of view. The other perspective is essentially objective—the "view from nowhere." Although objectivity is a major conceptual achievement, Nagel thinks it is not the whole story. There is more to reality than objective reality.

Could an objective science of subjectivity be possible? At times, Nagel seems to be optimistic about the possibility of an eventual theory. According to Nagel:

The strange truth seems to be that certain complex, biologically generated physical systems, of which each of us is an example, have rich nonphysical properties. An integrated theory of reality must account for this, and I believe that if and when it arrives, probably not for centuries, it will alter our conception of the universe as radically as anything has to date.

Newtonian mechanics did not include electricity or magnetism. Maxwell, Lorentz, and eventually Einstein showed how to incorporate those forces. Perhaps we are waiting for a Maxwell, a Lorentz, and an Einstein of consciousness. At times, Nagel seems to be pessimistic about any objective theory of subjective consciousness. Our inability to get an objective grasp on subjectivity may not be a historical limitation but the human predicament.

Can we solve the mind-body problem? Colin McGinn says no. McGinn first attempts to show something weaker: It is at least *possible* that the answer to the problem will be beyond our conceptual reach. McGinn says, "What is closed to the mind of a rat may be open to the mind of a monkey, and what is open to us may be closed to the monkey … Armadillo minds cannot solve problems of elementary arithmetic, but human minds can." Why think we have the conceptual keys to the entire universe? Why think all problems are within human reach?

McGinn then attempts to show that we *cannot* solve the mind-body problem. In order to answer the problem, we would have to have a conceptual link between subjective experience and physical reality. But our concepts of

subjective experience come from introspection. Those won't reach to the physical-reality side. Our concepts of physical reality come from perception of external objects. Those won't give us the subjective-experience side. None of our concepts, then, could possibly form a link between the two. McGinn's argument relies on questionable claims regarding the sources and limits of human concepts.

Antitheories say, "There is no answer to this question; give up." If we accept the antitheory and it is wrong, we will never get an answer … even if one exists.

Perhaps we should stack the deck against the antitheories. Science is more often a matter of how the evidence leans than of conclusive proof. Important in science, as in a court of law, is burden of proof. Some theories have particularly strong implications for scientific practice itself. Theories that say "The only possible answer is this kind of answer" imply that research must go in a certain direction. If we accept the theory and it is wrong, we will have stifled crucial research. Antitheories say, "There is no answer to this question; give up." If we accept the antitheory and it is wrong, we will never get an answer … even if one exists.

Two questions seem to be at issue:

- Might the antitheories be true?
- Should we believe the antitheories are true?

Although the answer to the first question may be yes, the answer to the second question is no. ■

Suggested Reading

Colin McGinn, "Can We Solve the Mind-Body Problem?" *Mind* 98.

Thomas Nagel, *The View from Nowhere*, introduction and Part I, "Mind."

David Rosenthal, "A Theory of Consciousness," *The Nature of Consciousness: Philosophical Debates*.

Questions to Consider

1. David Rosenthal says that a mental state is conscious when there is some "higher-order mental state" that is *about* it. What aspects of consciousness might that explain? What aspects might it not explain?

2. Are we ever justified in concluding that a scientific question is impossible to answer? Under what conditions are we justified in reaching such a conclusion?

What We Know and What We Don't Know

Lecture 24

I want to use this lecture to review in broad strokes some of the material that we've covered—some of the things we know—but also to emphasize some of the things that we don't know. The fundamental aim of science is a better grasp of reality. Its history is a history of repeated attempts and, of course, repeated failures.

The fundamental aim of science is a better grasp of reality. Its history is a history of repeated attempts and, of course, repeated failures. The history of science is a graveyard of theories that weren't quite good enough, in each case buried by theories we thought were better. In any inquiry, it is important to appreciate both the extent of our knowledge and the extent of our ignorance. This is especially important in any philosophical inquiry because philosophy focuses on questions that we don't quite yet know how to answer.

The history of science is a graveyard of theories that weren't quite good enough, in each case buried by theories we thought were better.

This lecture offers a review of some of the high points of the course, structured in terms of the examples with which we began: Descartes' dream, Einstein's brain, and Babbage's steam-driven computers. The course has reviewed and applied concepts from computer science in posing and trying to answer questions about minds. The lectures have examined Cartesian Dualism and its conceptual alternatives, offering a full survey of competing contemporary theories of consciousness and mind. In order to fully appreciate points of debate within and across disciplines, we have examined in detail a range of work in psychology and results in the brain sciences.

This final lecture tries to isolate the most important areas of our ignorance in terms of questions for further thought and further investigation. What is it that we *don't* know about Descartes' dream, about Einstein's brain, about Babbage's machine? What are our best options for answering those

questions? The most important questions about mind may be those we don't yet know how to ask.

The fundamental aim of science is a better grasp of reality. *Philosophy* means "love of wisdom." One might say that science aims at knowledge, while philosophy aims at wisdom. For both endeavors, it is important to have a grasp both of what we think we know and of those vast areas of ignorance that remain. Epistemology is the study of our sources of knowledge. Alasdair MacIntyre has suggested we need a parallel study of our sources of ignorance. This lecture offers a summary of some of the high points of the course but with an emphasis on what we do not know, as well as what we do. The introductory lecture began with three exhibits: Descartes' dream, Einstein's brain, and Babbage's steam-driven computers. In each category, we can catalog both knowledge and ignorance.

Let us start with Babbage's machine. What lessons do computing machines and artificial intelligence have for philosophy of mind? We have emphasized the connections between the history of logic and the history of computing. We have traced the work of Alan Turing and both traditions of artificial intelligence: GOFAI and the Connectionist alternative. We have traced the history of robotics from ancient myth to contemporary film. We have questioned whether genuine AI is possible. We have reviewed Turing's halting problem and Gödel's incompleteness theorem.

There is a great deal in this area that we do not know. We do not know where continuing attempts at artificial intelligence will lead. This issue echoes a controversy between Ada Lovelace and Turing regarding the possibility of creative machines. We do not know what attempts at artificial intelligence will reveal about our own brains. In a classic study, Terri Lewis found that infants

Although many once thought that infants' response to faces is hardwired, recent research has called that into question.

respond to images of faces shortly after birth. The traditional explanation of this result is that infants are hardwired with a preference for human faces.

More recently, researchers at the University of California at San Diego found that an infant-sized robot could quickly learn to recognize faces by following two simple rules: First, take note of unusual sights and sounds; second, focus on and link those events. With such experiments, work in artificial intelligence may tell us about our own intelligence. We have only begun to worry about the ethical consequences of building machines comparable to, or more intelligent than, ourselves.

Our second exhibit was Einstein's brain. Results and issues from the neurosciences have been a major resource throughout these lectures. The lectures have emphasized twin themes of localized brain function and brain plasticity. Brain function at the level of neurons and in terms of the overall organization of the brain has played a major role in our discussion of perception. Much of our discussion of brains has been functional rather than anatomical. Here, psychological and neuroscientific questions blend into conceptual philosophical questions regarding our concepts of mental function.

The history of the Molyneux problem illustrates this interaction. In 1688, Molyneux wrote John Locke a letter in which he asked whether a man who had been blind from birth but who had learned to distinguish shapes by touch would recognize those shapes by sight when his vision was restored.

- Locke claimed that he would not.

- Leibniz gave a different answer.

What do we know of Einstein's brain? What don't we know? We know a great deal about the brain at the lowest level and the highest level. We are abysmally ignorant about the vast area in between.

Even at the functional level, there is much that we do not know. An example from Diana Deutsch shows a connection between sound perceived as speech and as music. We don't yet understand that connection. What is

happening in synesthesia, in which the senses seem to overlap? Some people who experience synesthesia claim to "see" sounds, for example. The elephant in the room regarding our ignorance of the brain is the mystery of consciousness.

The lectures began with Descartes' dream. We have drawn from ancient, classical, and contemporary philosophy in pursuing issues of mind, using a wealth of thought experiments. The emphasis throughout has been on the force of argument in deciding between intellectual options and positions.

What do we know about Descartes' dream? What don't we know? Current philosophical consensus seems to be that we have left Descartes' dream behind, at least in the way he formulated it. Functionalism is the dominant theory in contemporary philosophy of mind, but it faces a major obstacle: the "hard problem of consciousness." For all our attempts to leave Cartesian Dualism behind, we still stand in its shadow.

Philosophical questions are often those we do not even yet know how to ask. Is the question of consciousness a scientific question, a technical question, or a conceptual one? Is it somehow all of these in one? ■

Suggested Reading

Rita Carter, *Mapping the Mind*, chapter 8.

John Vacca, ed., *The World's 20 Greatest Unsolved Problems*, chapter 14.

Questions to Consider

1. What do you think is the most important piece of information we have learned about minds and brains over the last 100 years?

2. What do you think is the most important area for exploration over the next 100 years?

Timeline

B.C.

c. 800 ..... Homer, *Iliad* and *Odyssey*.

c. 347–322 ..... Aristotle, *Prior Analytics*, *Posterior Analytics*, *De Interpretatione*.

c. 100 ..... The Antikythera mechanism, a Greek device used to calculate solar, lunar, and astrological positions. Found in 1902 from an ancient shipwreck.

A.D.

8 ..... Ovid, *Metamorphoses*.

1600–1800 ..... The golden age of automata.

1614 ..... John Napier invents logarithms.

1617 ..... Napier's bones.

1623 ..... William Schickard mechanizes Napier's bones.

1632–1633 ..... Galileo publishes *Dialogue Concerning Two World Systems* and is then tried for suspicion of heresy by the Catholic Church.

1629–1649 ..... René Descartes, *Meditations on First Philosophy*, *Passions of the Soul*, *De Mundo*.

1642–1662 ..... Blaise Pascal invents the Pascaline calculator and writes *Pensées*.

1651–1655 ..... Thomas Hobbes, *Leviathan* and *De Corpore*.

1666..... Isaac Newton, prism experiments on color.

1672–1715 ..... Gottfried Wilhelm Leibniz builds a multiplying machine.

1687..... Isaac Newton, *Philosophiae Naturalis Principia Mathematica*.

1689..... John Locke, *An Essay Concerning Human Understanding*.

c. 1737..... Jacques de Vaucanson builds his famous automatons.

1748..... David Hume, *An Enquiry Concerning Human Understanding*.

1748..... Julien Offray de La Mettrie, *Man a Machine*.

1769..... Wolfgang von Kempelen creates a fake chess-playing automaton.

1781..... Immanuel Kant, *The Critique of Pure Reason*.

1800–1821..... Francis Gall develops phrenology.

1801..... Jacquard's loom, an important step in the history of computing devices.

1810..... Johann Wolfgang von Goethe, *Theory of Colors*.

1820–1840..... Charles Babbage, Difference Engines #1 and #2 and the Analytical Engine.

1848..... An explosion drives a steel bar through Phineas Gage's head.

1854..... George Boole, "An Investigation of the Laws of Thought, on Which Are Founded the Mathematical Theories of Logic and Probabilities."

1859..... Charles Darwin, *On the Origin of Species*.

1861..... James Clerk Maxwell lays the foundations for color photography.

1864..... James Clerk Maxwell, "A Dynamic Theory of the Electromagnetic Field."

1872..... Charles Darwin, *Expression of the Emotions in Man and Animals*.

1874..... Franz Brentano, *Psychology from the Empirical Standpoint*; T. H. Huxley, "On the Hypothesis that Animals are Automata, and its History."

1890..... William James, *The Principles of Psychology*.

1891..... Sigmund Freud, *On Aphasia*.

1892..... Hendrik Lorentz develops the Lorentz field equations.

1904..... Alfred Binet develops an intelligence test that lays the foundation for IQ testing.

1910–1913 ..... Bertrand Russell and Alfred North Whitehead, *Principia Mathematica*.

1921..... Ludwig Wittgenstein, *Tractatus Logico-Philosophicus*.

1931..... Kurt Gödel, *On Formally Undecidable Propositions of Principia Mathematica and Related Systems*, Gödel's incompleteness theorem.

1937..... Alan M. Turing creates the Turing machine theory of computation.

1938..... B. F. Skinner, *The Behavior of Organisms*.

1943..... Jean-Paul Sartre, *Being and Nothingness*.

1946..... ENIAC and the von Neumann architecture for computing.

1949..... Gilbert Ryle, *The Concept of Mind*.

1950..... Alan M. Turing, "Computing Machinery and Intelligence."

1953..... James D. Watson and Francis Crick discover the molecular structure of DNA;

Ludwig Wittgenstein, *Philosophical Investigations*.

1956..... Dartmouth Artificial Intelligence Conference.

1958..... Frank Rosenblatt develops the perceptron, a two-layer, feed-forward neural net.

1962..... Thomas S. Kuhn, *The Structure of Scientific Revolutions*.

1965..... Hubert Dreyfus, "Alchemy and Artificial Intelligence."

1967..... Hilary Putnam, "The Nature of Mental States."

1969..... Marvin Minsky and Seymour Papert trash Rosenblatt's neural nets in perceptrons.

1974..... Thomas Nagel, "What Is It Like to Be a Bat?"

1979..... J. J. Gibson, *The Ecological Approach to Visual Perception*.

1980..... John Searle, "Minds, Brains, and Computers."

1981..... Paul Churchland, "Eliminative Materialism and the Propositional Attitudes."

1982..... Frank Jackson introduces black-and-white Mary in "Epiphenomenal Qualia"; Benjamin Libet's experiments on the

timing of readiness potential in the brain and consciousness of willing.

1987 ..... The resurrection of neural nets in Rumelhart and McClelland's *Parallel Distributed Processing*.

1988 ..... Daniel Dennett, "Quining Qualia"; Hans Moravec, *Mind Children: The Future of Human and Robot Intelligence*.

1989 ..... Roger Penrose, *The Emperor's New Mind*; Colin McGinn, "Can We Solve the Mind-Body Problem?"

1990 ..... David Rosenthal, "A Theory of Consciousness."

1991 ..... Rodney Brooks, "Intelligence without Representation"; Institution of the Loebner Prize.

1994 ..... Francis Crick, *The Astonishing Hypothesis*.

1995 ..... Patricia Smith Churchland, *Neurophilosophy*; Paul Churchland, *The Engine of Reason, the Seat of the Soul*; David Chalmers, "Facing Up to the Hard Problem of Consciousness."

1996 ..... Penrose and Hameroff's quantum theory of consciousness.

1997 ..... IBM's Deep Blue wins against chess grandmaster Gary Kasparov.

1998..... Andy Clark and David Chalmers, "The Extended Mind."

2000..... The date by which Alan M. Turing predicted in 1950 that a machine would pass the Turing test.

2040 ..... The date by which Ray Kurzweil predicts the "coming singularity," in which our machines will surpass us in intelligence.

Glossary

addiction transference: The phenomenon of replacing an addiction to one substance or action with an addiction to another substance or action, usually while attempting to cure the addiction to the first.

affordances: A concept developed in the work of J. J. Gibson, affordances are possibilities for action in an environment.

agnosia: Often the result of brain damage, agnosia is a general term for deficits in a person's ability to interpret information from the senses, despite a lack of damage to the senses themselves. See also **object blindness**, **prospagnosia**, and **visual agnosia**.

Analytical Behaviorism: The view that mental states can be analyzed or defined in terms of behavior or behavioral dispositions. See also **Behaviorism** and **Functionalism**.

Analytical Engine: Charles Babbage drew up plans for this general-purpose computing machine in the mid 1800s; though never built, the device was to be powered by steam and would have used programs on punch cards. See also **Difference Engine**.

Antikythera machine: Discovered just off the coast of the Greek island of Antikythera, the Antikythera machine was a Greek calculating device from about 100 B.C. that could have been used to predict the movement of the Sun, Moon, and major planets.

antinomies: Sets of compelling arguments on both sides of an irresolvable issue. In his *Critique of Pure Reason*, Kant presents the free will and Determinism debate as an antinomy.

artificial intelligence (AI): The science of making machines do things that would require intelligence if done by people. See **Marvin Minsky** in the Biographical Notes.

Asimov's Laws of Robotics: The robots featured in the fiction of writer Isaac Asimov must follow three laws: They cannot intentionally injure human beings; they must obey humans (when this does not conflict with the previous law); and they must protect themselves (again, when this does not conflict with the previous laws).

automata: Machines built to look like and imitate life forms—people or animals—with the appearance of autonomous action.

axiomatic systems: Systems of axioms or first principles from which other claims can be derived as theorems.

axon: Neurons, the cells of nerves, are equipped with a long protuberance called an *axon*. The signal of a neuron travels down the axon and is transmitted to other cells. See also **dendrites**, **neurons**, **neurotransmitters**, and **synapse**.

backpropagation of errors: A training process in which artificial neural nets (computer-instantiated structures loosely based on networks of neurons in the brain) learn from their mistakes.

Behaviorism: In psychology, a research program that seeks to understand the human mind in terms of behavioral inputs and outputs, that is, how humans react to different stimuli and changes in their environment.

binding problem: The binding problem refers to the question of how our brains assemble data from different aspects of perception into the unified consciousness that we experience.

blindsight: Despite their claim to be totally blind, some victims of major damage to the visual cortex are nonetheless able to "guess" about visual information in their environment with amazing accuracy. Paul Humphreys

speculates that blindsight functions by means of visual processing through an older and subconscious route.

body schema: Used by some researchers to indicate aspects of body image that are assumed unconsciously and through which one acts.

Boolean function: Any of various functions, familiar from truth tables that take truth values as input and give truth values as output. AND is a Boolean function, for example, which gives "true" as output only when both of its inputs are "true":

P	***Q***	***P* AND *Q***
T	T	T
T	F	F
F	T	F
F	F	F

brain state: A particular configuration of brain activities at a given moment. The term is often used in discussions regarding the philosophy of mind when considering whether mental states, such as belief or love, could reduce to physiological brain states.

Broca's and Wernicke's areas: Two parts of the brain associated with speech functions. Broca's area is involved with the production of speech. Wernicke's area is involved with the comprehension of words uttered by others.

C fibers: A category of slower-conducting nerve fibers responsible for less specific and slower-moving sensations, including pain.

Cartesian doubt: In his effort to determine whether there was anything of which he could be absolutely certain, René Descartes subjected all his knowledge to systematic doubt, rejecting anything about which he might possibly be mistaken or deceived. Descartes concludes that only the fact that he is doubting or thinking is beyond all doubt and, from that conclusion,

deduces that he exists. Descartes' *cogito, ergo sum*—"I think, therefore I am"—becomes the foundation of his system.

cognition: A general term for processes of conceptualizing, perceiving, and knowing.

cognitive psychology: In contrast to Behaviorism, a research program that attempts to understand the human mind in terms of the relation of inner mental states.

color circle, wheel, and solid: The color circle portrays colors in terms of similarities and oppositions. The color wheel adds a parameter of saturation, with the colors at the edges fully saturated and progressively mixed as one goes toward the center. The color solid is a three-dimensional portrayal that adds the parameter of intensity or brightness.

Compatibilist strategy: In the context of the Determinism argument, the strategy of challenging the assumption that free will and Determinism are necessarily opposed. The Compatibilist holds that free will, when properly understood, will be seen to be a natural part of a causal universe.

cones: Located on the retina, cones are one of the two types of light-receptive cells: rods and cones. Cone cells function in situations of normal light and register color. Different types of cones, using different pigments, specialize in different ranges of light wavelength. See also **optic nerve**, **retina**, and **rods**.

Connectionism: An interdisciplinary movement that attempts to explain the mind in terms of a large number of simple interconnected units, based on the fact that the brain is composed of an interconnected system of neurons. See also **backpropagation of errors**, **neural nets**, and **parallel distributed processing**.

Copernican theory: Nicolaus Copernicus's heliocentric theory in which the planets revolve around the Sun. See also **Ptolemaic astronomy**.

corpus callosum: The structure connecting the right and left hemispheres of the human brain that allows information to pass between the two. See also **split-brain patients**.

creature consciousness: As outlined by David Rosenthal, creature consciousness is contrasted with state consciousness and transitive consciousness as the state of a creature being awake or aware. Contrast **state consciousness** and **transitive consciousness**.

defeasible reasoning: A form of logic in which default assumptions operate until revised or qualified by new information.

deflationist response: A response that "deflates" a concept—of consciousness or freedom, for example—by showing that an opponent's assumptions regarding the concept are overinflated.

dendrites: Parts of neurons, dendrites serve as the receptors to the signals relayed by other neurons. See also **axon**, **neurons**, **neurotransmitters**, and **synapse**.

Determinism: In one sense, the claim that every event in the universe is the product of earlier events in accord with natural laws. In another sense, the claim that there can be no free will because all events are the product of earlier events. See also **free will**.

Difference Engine: A machine designed by Charles Babbage in the mid-1800s to calculate and print logarithm tables. Although not completed in his lifetime, Babbage's Difference Engine #2 was finally built in the 1990s by the London Science Museum.

dispositional properties: Properties (such as soluble or fragile) that reflect what would happen in certain circumstances. In Analytical or Philosophical Behaviorism, mental concepts are said to be synonymous with behavioral properties, generally understood as dispositional.

Dualism: In classical form, the position that the universe is composed of two radically different substances: the mental and the physical. For

Descartes, the mental does not occupy space as the physical does. See also **Epiphenomenalism**, **Monism**, **Occasionalism**, and **Parallelism**.

echolocation: A process of determining distance through the use of sound, used by such animals as bats and dolphins.

Einsteinian physics: Einstein's theory of relativity (general and special) challenged Newtonian tenets, holding, for example, that matter and energy are interchangeable, that time moves at a rate relative to one's rate of speed, and that space itself can be curved by gravity.

empirical: Deriving from experience of the world. *Scientific* is a rough synonym.

Empiricist theory: The theory that perception is a process of inference from sense-data. Classical Empiricists included Locke, Berkeley, and Hume, but the influence of Empiricism extended well into the 20th century in both philosophy and psychology.

Epiphenomenalism: The view that the mind does not have physical effects but merely "floats above" the physical processes of the brain. See also **Dualism**, **Occasionalism**, and **Parallelism**.

epistemology: That field of philosophy devoted to the study of knowledge and how we come to know things.

ethics: The field of philosophy that focuses on moral issues: ethically good actions, ethically right actions, rights, and obligations.

eugenics: The attempt to "advance" humanity by the selective breeding of human beings.

Evolutionary theory: In biology, the theory advanced by Charles Darwin that explains the development and complexity of species through the process of natural selection.

Existentialism: An influential movement in philosophy that took human freedom and one's capacity to create meaning for oneself as starting points.

faculty psychology: An approach to the human mind in terms of a limited number of powers or capacities (that is, faculties).

frame problem: An issue of relevance, the frame problem in artificial intelligence and in understanding human cognition is the problem of deciding what old information should be considered for revision in light of new information.

free will: The question of whether human beings should be considered to have free will, that is, whether they should be understood to have autonomous control over their own actions, is central to the history of philosophy. Answers given also bear on issues of moral and legal responsibility. See also **Determinism**.

frontal lobe: The foremost portion of the brain, understood to be an area important for planning and decision-making.

Functionalism: The position that mental states are functional states of an organism. Mental states, according to the Functionalist, take environmental input and other mental states as input, with behavior and other mental states as outputs.

ganglion cells: In visual perception, the cells that receive information from the retina and transport it to the brain. The long axons of these cells constitute the optic nerve.

Gödel's incompleteness theorem: Kurt Gödel proved that any axiomatic system adequate for simple arithmetic, if consistent, will be incapable of proving some truth expressible in the system.

GOFAI: Short for "good old-fashioned artificial intelligence," an attempt to produce artificial intelligence using rule-governed programs of symbol manipulation. Contrast with **Connectionism**.

halting problem: In the work of Alan M. Turing, the problem of deciding for any given program whether that program will "halt" as opposed to going into an infinite loop.

higher-order thought: Higher-order thought (HOT) theories attempt to analyze consciousness in terms of mental states that are about other mental states.

hippocampus: A seahorse-shaped part of the brain important in the procedures of processing emotions and producing memories.

holistic view (of the mind): The position that the phenomenon of mind is not the result of a specific region of the brain or a single process but emerges from the function of the brain as a unified whole.

homunculus: The "little man inside." In philosophical discussions of the "inner theater," the image of a homunculus is used to denigrate theories that would explain outer perception in terms of some form of inner perception. In brain structure, the term *sensory homunculus* is used to designate a model of the human body in which the proportional sizes of parts of the model correspond to proportional areas of representation in the sensorimotor cortex.

Idealism: Sometimes called *Subjective Idealism*, the response to the mind-body problem that holds that the physical world is illusion and only the mental realm exists. See also **Materialism**.

inference: In logic, the derivation of a conclusion from information contained in the premises.

inhibitory neurons: In contrast to the more common excitatory neurons, these nerve cells inhibit the firing of their target neuron, instructing it not to fire.

inner theater: In the philosophy of mind, the inner theater is used to designate an inner realm in which representations of the world are presented. See also **homunculus**.

intelligence quotient (IQ): The standard way of scoring intelligence tests. In classical form, a person's mental age (the age for which a person's test score is typical) was divided by his or her chronological age in years, with the result multiplied by 100. The average IQ is 100 by definition, with scores for the general population forming a normal or bell curve.

Intentionalist theory of perception: A theory emphasizing that perception is always "perception of" (Brentano), that perception comes with content, rather than content having to be added by inference. Contrast with **Empiricist theory**.

interaction problem: If the mental and the physical are two radically different realms, as Dualism claims, how could they possibly interact? This is the interaction problem posed for Dualism.

inverted lenses experiment: A form of psychological experiment in which a subject's visual stimuli are inverted or reversed by a lens. Results show that subjects accommodate to the reversal over time, though debate remains as to whether this is due to an inner perception that "flips" to accommodate or simply because they learn to work in terms of the new patterns of stimulation.

inverted spectrum: A thought experiment in which one person's color sensations are exactly opposite to another's; one person's qualitative sensations are blue where another's are yellow, for example.

Kurzweil's "coming singularity": Author and inventor Raymond Kurzweil foresees a near future in which our machines will become more intelligent than we are, capable of producing new machines still more intelligent than themselves.

limbic system: A set of structures deep within the brain involved with emotions and emotional memory.

Loebner Prize: A form of the Turing test run each year, offering $100,000 and a gold medal for the first computer program indistinguishable from a human. See http://www.loebner.net/ Prizef/loebner-prize.html.

logic: The study of patterns of valid deduction. Formal logic represents the essential structure of claims in symbolic form, codifying logical argument in the form of symbolic derivations. Mathematical logic studies formal properties of systems of logic. Philosophical logic concentrates on philosophical assumptions crucial to different logical systems.

Materialism: A response to the mind-body problem that holds that only the physical is ultimately real. *Reductive Materialism* claims that the realm of the mental somehow reduces to the physical. *Eliminative Materialism* claims that our concepts for the mental will be eliminated in an ultimately satisfactory and entirely physical scientific theory of human functioning. See also **Idealism**.

mental age: A controversial concept found in intelligence testing; used to indicate the age for which a person's score on the test would be typical.

mental set or set expectation: In perception, a background expectation that may influence what is perceived.

metaphysics: The most general conceptual investigation into the nature of reality.

mind-body problem: The mind-body problem refers to our conceptual difficulty in understanding the relation between mental and physical phenomena. For different answers to the problem, see **Dualism**, **Epiphenomenalism**, **Functionalism**, **Idealism**, **Materialism**, **Occasionalism**, and **Parallelism**.

Monism: In the philosophy of mind, the position that there exists only one basic kind of "stuff" or substance. Both Materialism and Idealism are forms of Monism, as opposed to Dualism, which holds that the universe contains two fundamentally different kinds of things: the realms of the mental and the physical. See also **Dualism**.

multiple instantiation: The notion that mental states might be instantiated in any of various forms of organisms and even, perhaps, in machines.

Naïve Realism: The view that the world as we perceive it is essentially the way the world actually is independent of our perception of it.

NAND: A Boolean connective that means "not both are true." The following is a truth table for NAND:

P	***Q***	***P* NAND *Q***
T	T	F
T	F	T
F	T	T
F	F	T

neural nets: Computational structures instantiated in software but modeled roughly on the operation of neurons in the brain. Trained by backpropagation of errors, neural nets have shown an impressive ability to generalize, that is, to learn patterns applicable to new cases. See also **backpropagation of errors**, **Connectionism**, and **parallel distributed processing**.

neurons: The cells of the nervous system. Stimulated by input at their dendrites, neurons pass a signal down their axons to their terminal nodes, where electrochemicals called *neurotransmitters* are released into a synapse and stimulate other neurons in turn. See also **axon**, **dendrites**, **neurotransmitters**, and **synapse**.

neurotransmitters: When one neuron transmits a signal to another, it does so by releasing chemicals called *neurotransmitters* into the small space (or synapse) between the neurons. The neurotransmitters cause a change in the receiving terminal (or dendrite) of the neuron being given the signal. See also **axon**, **dendrites**, **neurons**, and **synapse**.

Newtonian astronomy: Isaac Newton's laws of gravitation codified and explained the movement of celestial bodies in accord with the Copernican (heliocentric) conception of planetary motion. But Newtonian physics also left unsolved a number of problems that paved the way for Einsteinian physics as a replacement.

object blindness: A specific type of visual agnosia characterized by the inability to distinguish the identity of objects.

Occasionalism: A form of Dualism, Occasionalism holds that the mental and physical are, in fact, causally isolated but operate in sync through the action of God at every moment. See also **Dualism**, **Epiphenomenalism**, and **Parallelism**.

optic nerve: The bundle of nerve fibers that brings visual information from the retina (the light-sensitive layer on the rear interior surface of the eye) to the brain.

optical illusion: An image or object that plays upon human processes, leading a person to misperceive what he or she sees. See examples in Lecture Ten.

Panpsychism: A response to the "hard problem of consciousness" defended by David Chalmers and Galen Strawson. The Panpsychist holds that all matter is in some way conscious.

paradigms: In the philosophy of science of Thomas S. Kuhn, a set of background assumptions and explanatory concepts definitive of a science at a particular time.

parallel distributed processing: A term associated with Connectionist neural networks. It involves systematically constructing networked lines of connections between units, which are strengthened and weakened depending on the successes or failures of the processes. See also **backpropagation of errors**, **Connectionism**, and **neural nets**.

Parallelism: A form of Dualism asserting that the mental and physical are, in fact, causally isolated but operate in "pre-established harmony" because both realms were wound up like two clocks by God at the inception of the universe. See also **Dualism**, **Epiphenomenalism**, and **Occasionalism**.

perception: The process of gaining awareness of something through one's bodily senses.

perceptron: A two-layer neural net developed by Frank Rosenblatt. Perceptrons could be trained using a simple rule to any logical function they could instantiate, but Minsky and Papert showed that perceptrons could not instantiate some simple logical functions, such as *exclusive or*.

phantom limb: The name for the experience of one who has lost an appendage yet has an illusory sensation of its presence.

phenomenology: A tradition in philosophy that takes subjective experience as its starting point. Objectivity and science, in this view, are seen as a second level of abstraction; one must step out of one's experience in attempting to acquire an objective perspective.

philosophy of language: The branch or topic of philosophy concerned with understanding how language is structured and used. Questions arise regarding the relationship between language and reality and how linguistic meaning should be understood.

philosophy of mind: The branch of philosophy concerned with understanding the nature of the mind, the nature of consciousness, and the relationship between minds and brains, or the mental and the physical. Contemporary philosophy of mind is aggressively interdisciplinary, interfacing with psychology, computer science, and the neurosciences.

phrenology: A pseudoscience popular and influential in the 1800s, which studied the bumps of a person's skull to learn about his or her character traits.

plasticity: Also called *neuroplasticity*, *brain plasticity*, and *cortical plasticity*, the ability of brain matter to alter so as to perform different functions. In learning new manual skills, areas of the brain may be recruited to new tasks.

pre-philosophical facts: Commonsense assumptions or understandings in advance of critical reflection.

private language argument: Ludwig Wittgenstein's *Philosophical Investigations* is structured as an assortment of separate comments, seen

by many as containing a central argument regarding the essentially public nature of language. In these lectures, the argument is presented in terms of a necessity for public criteria in language learning and, thus, language comprehension, though there is much disagreement as to precisely what the argument is and how or whether it works.

privileged access: The notion that we each have access to the contents of our own minds in ways that others do not.

prosopagnosia: Often the result of damage to the brain, prosopagnosia is a characterized by the inability to recognize faces, despite being able to recognize other objects without difficulty.

Ptolemaic astronomy: Ptolemy's geocentric (that is, Earth-centered) model of the universe was the dominant theory for hundreds of years, rigorously articulated but eventually superseded by Copernicus's heliocentric model. See also **Copernican theory**.

qualia: Plural for *quale*, *qualia* refers to subjective qualitative experiences, for example, the taste of a pineapple or the feel of silk.

quantum mechanics: Developed early in the 20th century, quantum mechanics is a sophisticated theory regarding subatomic events. Although the theory is well confirmed experimentally, its interpretation remains an area of controversy. Its implications or claimed implications extend to whether Determinism is true, whether every event has a cause, whether conscious measurement is constitutive of the universe, and even what the nature of human freedom might be.

quantum randomness: In standard interpretations of quantum mechanics, events occur at the quantum level that have no specific cause and are impossible to predict.

Reductive Materialism: That form of Materialism that holds that the mental can be reduced to the physical. In terms of the relation between the things at issue, Reductive Materialism claims that mental things are ultimately purely physical. In terms of the sciences at issue, Reductive Materialism

claims that the science of the mental will follow directly from the science of the physical.

relational properties: A property something has in virtue of its relation or interaction with another thing. "Beside" and "married" are relational properties: One thing cannot be "beside" all by itself, nor can a person be "married" without another person. Relational properties are contrasted with intrinsic properties, which belong to something independent of their relations to others.

retina: The layer of light-sensitive cells that covers the rear interior of the eye. The retina registers incoming light as the beginning of a process of signals sent to the brain. See also **cones**, **optic nerve**, and **rods**.

robotics: The study and design of robots, machines that work somewhat autonomously to perform tasks for humans.

rods: Located on the retina, rods are one of the two types of light-receptive cells: rods and cones. They function in situations of little light and do not register color. See also **cones**, **optic nerve**, and **retina**.

saccades: Our eyes move in swift jumps called *saccades* (or *saccadic motion*) as we scan across something, such as the page of a book or a computer screen.

semantics: The meaning of the symbols of a system of language or the study thereof. Semantics is contrasted with syntax, a matter of the shapes of the symbols and rules in terms of those shapes. Semantics concerns the relation of those shapes to ideas and things in the world. See also **syntax**.

sense-data: In the Empiricist theory, information delivered to the brain via the senses that is then used as the basis for the inferences that form our experience of the world.

sensory cortex: That part of the cortex that registers touch from various parts of the body; the sensory cortex is organized in a way that corresponds roughly to the organization of the body. See also **homunculus**.

Solipsism: The position that the only thing that exists is one's own mind.

spandrels: In Evolutionary theory, properties of organisms that were not directly selected for but nonetheless "came along for the ride."

split-brain patients: People in whom the corpus callosum between the two hemispheres of the brain has been surgically cut, often to relieve extreme epilepsy. Split-brain patients function normally in most contexts but show surprising behavior in carefully constructed experimental conditions because of the lack of communication between the hemispheres.

state consciousness: In the work of David Rosenthal, the sense in which a mental state is conscious—for example, a belief is a conscious belief or anger is conscious anger. Contrast **creature consciousness** and **transitive consciousness**.

Stoics: A school of ancient philosophy known for its work in logic, physics, and ethics. The Stoics held that Determinism was true and that we have no control over the "slings and arrows of outrageous fortune." The rational road to tranquility is to control one's emotional reactions to the inevitable.

strong AI: The position that a machine can have a mental state, such as understanding, in virtue of instantiation of a particular program. Weak AI is a research program in which programs are used to understand mental states. Strong AI is the claim that an appropriate program would be a mental state.

Substance Dualism: Also known as *Cartesian Dualism*, the position that the universe is composed of two radically different kinds of "stuff" or substances: the realm of the mental and the physical. See also **Dualism**, **Epiphenomenalism**, **Occasionalism**, and **Parallelism**.

symbol-processing: The methodical manipulation of symbols. A calculator does math by manipulating symbols according to rules, for example, rather than by dealing with the quantities the numbers may represent.

synapse: In sending a signal to another neuron, the neuron ejects neurotransmitters into the small space between the two neurons. That space is called a *synapse*. See also **axon**, **dendrites**, **neurons**, and **neurotransmitters**.

syntax: The structure of the symbols in a language or the study thereof. Syntax is a matter of the shapes of the symbols and rules in terms of those shapes. Semantics, in contrast, concerns the relation of those shapes to ideas and things in the world. See also **semantics**.

transitive consciousness: In the work of David Rosenthal, transitive consciousness is consciousness of something. Contrast **creature consciousness** and **state consciousness**.

trichromacy: The normal human capacity to see colors. The term refers to the fact that color perception functions in terms of three sets of color-sensitive cones in our retinas. See also **cones** and **retina**.

Turing machine: An abstract machine conceptualized by Alan M. Turing as a formal model for the concept of computation. The Turing machine served as a model for the building of real computers.

Turing test: Alan M. Turing suggested that the question "Can a machine think?" be replaced with a specific test: In communication through a monitor interface, can a computer fool a person into thinking that it, too, is a person?

unity of consciousness: Despite receiving a variety of sensory data, processed in various areas in the brain, we experience consciousness as a seamless unity.

valid: An argument is valid if the conclusion follows from the premises. A deductively valid argument is one in which the connection is logically tight and in which it is logically impossible for the premises to be true and the conclusion to be false.

visual agnosia: A form of agnosia that results in an inability to interpret visual data. Both object blindness and prosopagnosia fall under this subcategory. See also **agnosia**, **object blindness**, and **prosopagnosia**.

visual cortex: Usually used to refer to the primary visual cortex located in the rear of the brain, which processes incoming visual data from the eyes.

volitional (or willing) dysfunction: A disorder that inhibits a person's ability to control his or her own actions.

voluntary action: An act made as the result of one's choice, contrasted with involuntary or reflex actions over which one does not have conscious control.

von Neumann architecture: An overall design for computers presented by John von Neumann in 1945 on the basis of work on the ENIAC. Virtually all contemporary computers have a von Neumann architecture, in which memory functions to contain both data and the program that operates on those data.

weak AI: Weak AI is a research program in which programs are used to understand mental states. Strong AI, in contrast, is the claim that an appropriate program would be a mental state.

zombies: In thought experiments in the philosophy of mind, zombies are supposed to be behaviorally and functionally identical to people but without consciousness or inner experience.

Biographical Notes

Aristotle (c. 384–322 B.C.): A major figure in Western philosophy and science, student of Plato, and teacher of Alexander the Great. His work covers topics in physics, poetics, rhetoric, ethics, epistemology, and metaphysics.

Antoine Arnauld (1612–1694): A French philosopher and theologian, Arnauld is best known for his adaptation and advancement of Descartes' philosophy.

Charles Babbage (1791–1871): A mathematician and mechanical engineer, Babbage is known for pioneering work in the development of computers. Babbage devised plans for Difference Engines #1 and #2, complex, steam-driven calculating machines designed to compute and print logarithm tables. His greatest design was for the Analytical Engine, which would have been a universal programmable computer.

Alfred Binet (1857–1911): In 1904, in an attempt to tailor elementary education to student needs, Binet developed the first modern intelligence test, the foundation for all later IQ testing. Binet's warnings that the test was not intended to measure some single quality called *intelligence*, that it should not be used to rank normal children, and that there was no reason to believe that whatever it measured was innate, were largely ignored.

Ned Block (b. 1942): Block is a professor of philosophy at New York University with research areas in philosophy of mind and cognitive science. He is known for the Chinese nation thought experiment.

Josh Bongard (b. 1974): Bongard, currently teaching at the University of Vermont, is a computer scientist working in robotics. He has built a robot that has the ability to learn its own body.

George Boole (1815–1864): A British mathematician, logician, and philosopher, Boole authored *Laws of Thought* and created what is now called *Boolean algebra*.

Cynthia Breazeal (b. 1967): A professor at MIT, Breazeal is known for the exploration of emotion in robotics. Her most complex robot, Leonardo, can read and respond to emotional cues in interaction with humans.

Franz Brentano (1838–1917): According to Brentano, "All perception is perception of." He is known for groundbreaking work in the philosophy of psychology and for the development of the concept of intentionality.

Rodney Brooks (b. 1954): A professor of robotics at MIT, Brooks has argued against the traditional symbolic manipulation approach to artificial intelligence in favor of an interactive and embodied one. Brooks calls for the progressive development of increasingly complex robots in interaction with the world.

Jerome Bruner (b. 1915): An American psychologist, Bruner's research explored dimensions of cognitive psychology and cognitive learning.

David Chalmers (b. 1966): Chalmers is a prominent figure in contemporary philosophy of mind, currently teaching at the Australian National University. He is best known for pressing the "hard problem of consciousness" and the "explanatory gap": How can subjective experience possibly be explained in terms of any physical substrate or functional organization?

Patricia Smith Churchland (b. 1943): Churchland, currently teaching at the University of California at San Diego, is a philosopher working in close contact with neurophysiology.

Paul Churchland (b. 1942): A philosopher of mind currently teaching at the University of California at San Diego and most well known for his doctrine of Eliminative Materialism. This doctrine proposes that our folk psychological concepts for mental states—belief, love, and consciousness, for example—will simply disappear with the development of an adequate

science of humans and their functioning in much the same way that concepts of witches and humors have been abandoned.

Andy Clark (b. 1957): Clark is a professor of philosophy at the University of Edinburgh. Specializing in philosophy of mind, Clark is known for work with David Chalmers on the extended-mind theory. Clark and Chalmers propose that a broad range of our mental activities—memory, belief, and thinking itself—are often partially constituted by things in the world beyond our skins.

Garrison Cottrell (b. 1950): Cottrell, a professor at the University of California at San Diego, is a computer scientist working in cognitive science. One of his achievements is a three-layer neural network successful in learning face recognition.

Francis Crick (1916–2004): Crick, an English molecular biologist, was co-discoverer, with James D. Watson, of the structure of the DNA molecule. In later work, Crick dedicated himself to neuroscience and the study of consciousness.

Antonio Damasio (b. 1944): A behavioral neurobiologist at the University of Southern California who has made important contributions to the interdisciplinary discussion between philosophy and the neurosciences. Damasio's main interest is the neurological systems involved in memory, emotions, and decision-making.

Charles Darwin (1809–1882): The central figure in evolutionary biology and one of the outstanding scientists of all time. Darwin's *On the Origin of Species*, the product of extensive research, provided incontrovertible evidence for the theory that all species evolved through time by a process Darwin called *natural selection*.

René Descartes (1596–1650): Descartes, a major figure in Western philosophy, was also an eminent mathematician and scientist. He is known for both Cartesian coordinates and Cartesian Dualism: the theory that the mental and the physical are two radically distinct aspects of the universe.

Diana Deutsch (b. 1938): Currently teaching at the University of California at San Diego, Deutsch is a perceptual and cognitive psychologist known for pioneering research in auditory illusions and the psychology of music.

Hubert Dreyfus (b. 1929): Dreyfus is a professor of philosophy at the University of California at Berkeley, known for his attacks on the prospects of artificial intelligence.

Euclid (c. 300 B.C.) Euclidean geometry is the historical paradigm of an axiomatically developed system, long thought to be the only possible view of relations in space. Over the last 200 years, alternatives known as *non-Euclidean geometries* have been developed.

Jerry Fodor (b. 1935): Fodor is a philosopher and cognitive scientist currently teaching at Rutgers University and known for his theory of the language of thought. The theory claims that thinking employs a mental language of representations, sometimes called *mentalese*.

Harry Frankfurt (b. 1929): Frankfurt, professor emeritus of philosophy at Princeton University, proposed insightful theories about free will. He argued for a hierarchical understanding of free will, in which free action is not just the ability to act on a desire but also includes a second-order volition that desires what to desire. An act is free if it is in accord with the desire one wants to desire.

Gottlob Frege (1848–1925): A German mathematician, logician, and philosopher, Frege is considered one of the founders of contemporary logic and philosophy of language. His attempt to ground all of mathematics on basic logical axioms inspired Russell and Whitehead's later work in *Principia Mathematica*.

Sigmund Freud (1856–1939): Freud, an Austrian-born psychiatrist and neurologist, founded psychoanalysis.

Phineas Gage (1823–1860): In a railroad construction accident, an iron rod was blasted through Gage's skull. He did not die but underwent a radical change in personality. Although used at the time as an argument for a holistic

approach to mind, Gage's story has become a classic example of functional localization in the brain.

Francis Gall (1758–1828): An Austrian anatomist, Gall was the founder of phrenology. Phrenologists thought they could detect the developed areas of the brain by feeling for bumps and valleys on the skull. Although phrenology is now considered a pseudoscience, the idea of functional localization is a basic tenet of contemporary neuroscience.

Gordon Gallup (b. 1941): Currently teaching at the University of Albany, Gallup is a psychologist with a specialty in biopsychology. He is known for the development of the mirror test, used to determine self-awareness in animals.

Howard Gardner (b. 1943): Gardner, currently teaching at Harvard University, is a psychologist well known for his theory of multiple intelligences.

J. J. Gibson (1904–1979): Gibson was an American psychologist with a specialty in the field of visual perception. In *The Perception of the Visual World*, he develops the theory of "affordances." In Gibson's theory, what we perceive are not objects but affordances—possibilities for action in the environment.

Henry H. Goddard (1866–1957): Goddard, an American psychologist and eugenicist, was the first to bring IQ testing to America. In violation of Binet's warnings, Goddard used the test as a ranking, assigning the terms *idiots*, *imbeciles*, and *morons* to those in the lowest three categories.

Kurt Gödel (1906–1978): Gödel was an Austrian American mathematician, logician, and philosopher. His incompleteness theorem had a profound effect on the philosophy of logic and mathematics in the 20th century. Gödel's theorem proves that for any axiomatic system that includes arithmetic, there will be truths about numbers that cannot be proven in the system.

Johann Wolfgang von Goethe (1749–1832): Although also a scientist and theologian, Goethe is best known as a key figure in German literature. His

Theory of Colors argued, against Newton, that color exists not in wavelengths of light but in the mind.

Stuart Hameroff (b. 1947): Hameroff is an anesthesiologist and professor at the University of Arizona, known for his research collaboration with Roger Penrose. Penrose and Hameroff suggest that the key to consciousness may be found in quantum effects in the microtubules of neurons.

N. R. Hanson (1924–1967): Hanson, an American philosopher of science, spent much of his philosophical energy exploring how observation is influenced by beliefs, making observation theory-laden. In *Patterns of Observation*, he argues that two people with different beliefs will experience the world in radically different ways.

Donald O. Hebb (1904–1985): An American psychologist, Hebb was the first to show how artificial networks of neuron-like devices could learn.

Thomas Hobbes (1588–1679): Although he is best known for his work in political theory in *Leviathan*, Hobbes's views on personal identity in *De Corpore* were also influential.

David Hume (1711–1776): Hume was an influential thinker in the Scottish Enlightenment and is considered one of the giants of British Empiricism. He also outlined an early theory of learning by association.

Nicholas Humphrey (b. 1943): A British psychologist and philosopher, Humphrey has spent of much of his professional life exploring issues of consciousness. In work on monkeys, he was the first to discover a mode of vision called *blindsight*, in which an individual is able to see in a certain sense despite complete damage to the visual cortex.

Edmund Husserl (1859–1938): A German philosopher, Husserl is the father of phenomenology. This philosophical method attempts to engage in a science of consciousness with the purpose of discovering the structure of experience.

Thomas H. Huxley (1825–1895): Huxley was a strong and early advocate of the theory of evolution. He is also associated with Epiphenomenalism, the view that mental states merely "float above" the physical states of the brain.

Frank Jackson (b. 1943): A philosopher at the Australian National University, Jackson has done extensive research in the philosophy of mind, metaphysics, and epistemology. He is best known for his "black-and-white Mary" argument that there are truths of consciousness that will forever escape physical science.

William James (1842–1910): A professor of philosophy at Harvard, James was an original thinker in the fields of both philosophy and psychology. James was a principal proponent of Pragmatism and did much to establish psychology as an empirical science.

Immanuel Kant (1724–1804): Kant is widely considered one of the giants of the European philosophical tradition. At the age of 57, he published his masterpiece, *The Critique of Pure Reason*.

Johannes Kepler (1571–1630): An influential German mathematician and astronomer, Kepler proved in *Astronomia Nova* that the Earth revolves around the Sun in elliptical orbits, which is his principal contribution to science. In communicating Napier's work to William Schickard, Kepler also played a role in the history of calculating machines.

Christof Koch (b. 1956): Koch is a neuroscientist at the California Institute of Technology. In collaboration with Francis Crick, he developed a theory of consciousness in terms of synchronized 40-Hertz firing of neurons in the brain.

Thomas S. Kuhn (1922–1996): A historian and philosopher of science who earned renown with *The Structure of Scientific Revolutions*. Kuhn claimed that the history of science represents not a progressive accumulation of new knowledge but the repeated and revolutionary overthrow of earlier scientific paradigms.

Ray Kurzweil (b. 1948): Kurzweil has made formative contributions in optical character and speech recognition, synthesized music, and speech production. Kurzweil warns of the "coming singularity," a point at which our machines will be more intelligent than we are.

Julien Offray de La Mettrie (1709–1751): A French physician and philosopher who argued in *Man a Machine* that animals were machines and that man is an animal. Man, therefore, is a machine.

Gottfried Wilhelm Leibniz (1646–1716): Leibniz was an influential philosopher, theologian, diplomat, physicist, and mathematician. He developed binary notation and built a complex calculating machine, and his development of calculus paralleled Newton's. He is known for the "indiscernibility of identicals," the principle that if two things are identical, anything true of one will be true of the other. In response to the interaction problems of Dualism, Leibniz proposed Parallelism. According to Parallelism, mind and body are not causally connected but function in parallel, like identical clocks wound up at the same time, because of a "pre-established harmony" initiated by God at the creation of the universe.

Benjamin Libet (b. 1916): Libet, an American physiologist, is an innovator in the science of human consciousness. Libet is known for experiments indicating that the brain's initiation of action often *precedes* conscious decision.

John Locke (1632–1704): A major figure in British Empiricism, Locke also developed a classical position regarding personal identity and memory. His work in political philosophy was highly influential on the American Declaration of Independence.

Elizabeth Loftus (b. 1944): A professor of psychology and law at the University of California at Irvine, Loftus's research has made her one of the world's leading experts on the fallibility of eyewitness testimony.

Hendrik Lorentz (1853–1928): Lorentz won the Nobel Prize in physics in 1902 for his work on electrodynamics and relativity. Unlike Maxwell, he did

not think magnetism could be explained in terms of classical mechanics and, in that way, paved the way for Einstein.

Nicolas de Malebranche (1638–1715): Malebranche tried to address the interaction of mind and body by championing a theory called *Occasionalism*. Mind and body are not causally connected; according to Occasionalism, God is involved at every moment in making them act in parallel.

James Clerk Maxwell (1831–1879): A Scottish mathematician and physicist, Maxwell is best known for the Maxwell equations, which first linked basic laws of electricity and magnetism.

Jay McClelland (b. 1948): McClelland, an American psychologist, is known for his work with artificial neural nets. In 1986, McClelland and David Rumelhart authored *Parallel Distributed Processing*, which introduced backpropagation of errors as a new learning rule for a new kind of net. Their work resurrected Connectionism.

Warren McCulloch (1898–1968): McCulloch was an American neurophysiologist. With Walter Pitts, he demonstrated in the 1940s how simple electrical devices could imitate neural firing.

Colin McGinn (b. 1950): McGinn is a professor at the University of Miami and a *mysterion*: In "Can We Solve the Mind-Body Problem?" McGinn argues that some things are forever beyond the range of human knowledge. One of these is an understanding of consciousness.

John Stuart Mill (1806–1873): A major contributor to the philosophical fields of logic, ethics, and political theory, Mill outlined an early theory of learning by association.

Marvin Minsky (b. 1927): Minsky is a major figure in the theory of computation and a pioneer in artificial intelligence. Shortly after the Dartmouth conference of 1956, he built a program capable of constructing proofs in geometry.

William Molyneux (1656–1698): Molyneux, an Irish-born scientist and politician, posed a question to John Locke that has become known as the *Molyneux problem*. If a man was born blind and learned to distinguish between basic geometric shapes by touch—a cube and sphere, for example—would he be able to distinguish them by sight once his vision was restored?

Hans Moravec (b. 1948): Moravec is professor at the Robotics Institute at Carnegie Mellon University. He lauds robotics as the next step in evolution.

Samuel George Morton (1799–1851): An American physician and natural scientist, Morton collected hundreds of skulls in the attempt to measure the intelligence of different races by comparing brain size.

Thomas Nagel (b. 1937): Nagel teaches at New York University. He has published numerous essays and books, the most important of which may be his essay entitled "What Is It Like to Be a Bat?" Nagel argues that the subjective quality of consciousness cannot be explained through objective science.

John Napier (1550–1617): Napier was a Scottish mathematician and physicist known for the development of logarithms and of Napier's bones, a calculating device for doing multiplication by addition alone.

John von Neumann (1903–1957): Hungarian by birth, von Neumann was a major figure in 20th-century physics, mathematics, and computer science. He is known for the von Neumann architecture that characterizes all contemporary computers, in which memory functions to contain both data and the program that operates on that data.

Allen Newell (1927–1992): Newell was a pioneer in the fields of computer science and artificial intelligence. He and Herbert Simon achieved one of the first major successes in artificial intelligence by creating the Logical Theorist, a program capable of proving theorems in formal logic.

Isaac Newton (1643–1727): Newton, calling himself a natural philosopher, is a paramount figure in the history of science. His 1666 prism experiments laid the foundations for contemporary optics and the theory of color.

Alva Noë (b. 1964): A professor of philosophy at the University of California at Berkeley, Noë specializes in philosophy of mind. In the tradition of J. J. Gibson, Noë argues for an enactment theory of perception: that perceiving is a way of acting.

Stephen Palmer (b. 1948): Palmer is a professor of psychology at the University of California at Berkeley, where he is the director of the university's Visual Perception Lab. Palmer's theoretical work has applied neuroscience to the philosophical issue of the inverted spectrum.

Seymour Papert (b. 1928): Papert, a professor at MIT, is a computer scientist, mathematician, and a pioneer of artificial intelligence. With Marvin Minsky, he launched a devastating attack on Rosenblatt's neural network perceptrons, rendering Connectionism obsolete until the mid-1980s.

Derek Parfit (b. 1942): Parfit, a British philosopher currently teaching at Oxford, specializes in issues of self-identity, rationality, and ethics and the relations among the three.

Blaise Pascal (1623–1662): Pascal was an eminent French mathematician, philosopher, and theologian. The inventor of the Pascaline, a calculating machine, he reacted to Descartes' claim that animals were merely unfeeling machines by saying, "I cannot forgive Descartes."

Roger Penrose (b. 1931): Penrose, professor at the University of Oxford, is a theoretical physicist and mathematician. Penrose has proposed that some aspects of human intelligence are non-algorithmic. Working with Stuart Hameroff, he suggests that the key to consciousness may be found in quantum effects in the microtubules of neurons.

Walter Pitts (1923–1969): A genius in logic and mathematics. Pitts never attended college but frequented lectures at the University of Chicago. With Warren McCulloch, he demonstrated in the 1940s how simple electrical devices could imitate neural firing.

Plato (c. 427–347 B.C.): A student of Socrates as well as the teacher of Aristotle. Plato's work is a formative part in the history of all major philosophical fields of study.

Hilary Putnam (b. 1926): An American philosopher central to 20th-century philosophy of mind, philosophy of science, and philosophy of language. He is known for the development of Functionalism, particularly in terms of the model of Turing machines. Functionalism identifies mental states with functional states of the organism.

Pythagoras (c. 580–c. 500 B.C.): Pythagoras and the Pythagoreans made major advancements in mathematics, including the Pythagorean theorem and the existence of irrational numbers.

V. S. Ramachandran (b. 1951): Born in India, Ramachandran is a neurologist and professor at the University of California at San Diego. He is best known for research that analyzes the phenomenon of phantom limbs in terms of body image in the brain.

Frank Rosenblatt (1928–1969): Rosenblatt developed the perceptron, a two-layer, feed-forward artificial neural net in the early 1960s. Using Rosenblatt's delta learning rule, it was shown that a perceptron could be trained to any pattern of input-output responses it could instantiate. Unfortunately, Marvin Minsky and Seymour Papert showed that some patterns existed that perceptrons could not instantiate, including *exclusive or*.

David Rosenthal (b. 1939): Rosenthal is professor at the City University of New York who focuses on philosophy of mind. He is known for his higher-order thought (HOT) theory of consciousness, according to which a mental state is conscious if it is the target of a higher-order thought.

David Rumelhart (b. 1942): Rumelhart, an American psychologist, is known for his work with artificial neural nets. In 1986, Rumelhart and Jay McClelland authored *Parallel Distributed Processing*, which introduced backpropagation of errors as a new learning rule for a new kind of net. Their work resurrected Connectionism.

Bertrand Russell (1872–1970): Russell was a significant figure in 20th-century logic and philosophy, with major contributions in the philosophy of logic, philosophy of language, philosophy of mind, and epistemology. Russell and Alfred North Whitehead's *Principia Mathematica* showed that all of mathematics could be built from a few simple logical concepts and, thereby, paved the road for contemporary computing.

Gilbert Ryle (1900–1976): Ryle was a follower of Wittgenstein and an influential proponent of Analytical Behaviorism. In *The Concept of Mind*, he attacks Cartesian Dualism, ridiculing the notion that the mind is some separate entity that inhabits a body as the dogma of the "ghost in the machine."

Oliver Sacks (b. 1933): Sacks is neurologist and writer known for a number of popular and influential books. *The Man Who Mistook His Wife for a Hat* and *An Anthropologist on Mars* investigate brain disorders and their relation to consciousness and cognition.

Jean-Paul Sartre (1905–1980): A French playwright, novelist, and philosopher, Sartre built the idea of freedom into the core of his Existentialism.

Roger Schank (b. 1946): With colleagues at Yale, Schank designed a program to address the frame problem in the understanding of narratives. Enthusiasts claimed that programs like Shank's could understand the story, a claim that stimulated John Searle's Chinese room thought experiment as a rebuttal.

William Schickard (1592–1635): On the basis of Napier's work and in communication with Kepler, Schickard developed an early calculating machine.

John Searle (b. 1932): A leading philosopher of mind, Searle is professor of philosophy at the University of California at Berkeley. He is famous for his Chinese room thought experiment, intended as a criticism of strong AI. Searle's example attempts to show that no machine, in virtue of instantiating a program, could be said to understand English or have other cognitive abilities.

Herbert Simon (1916–2001): Simon was a political scientist and a pioneer in the fields of computer science and artificial intelligence. He and Allen Newell achieved one of the first major successes in artificial intelligence by creating the Logical Theorist, a program capable of proving theorems in formal logic.

B. F. Skinner (1904–1990): An American psychologist, Skinner is known as a foremost figure in psychological Behaviorism and, in particular, for the development of schedules of reinforcement.

Roger Sperry (1913–1994): Sperry won the Nobel Prize for his work with split-brain research. In treating epileptics, he severed the corpus callosum, the part of the brain that connects the left and right hemispheres of the brain. According to Sperry, "Everything we have seen indicates that the surgery has left these people with two separate minds … that is, two separate spheres of consciousness."

Thales (c. 624–c. 546 B.C.): Thales is often called the first Greek philosopher and scientist. In contrast to mythological explanations of the world, Thales attempted to devise a natural and rational one.

Louis L. Thurstone (1887–1955): Thurstone, an American psychologist, is known for his work on intelligence testing. He argued against the notion of a general intelligence, suggesting that individuals could possess a number of "primary mental abilities" in varying degrees.

Alan M. Turing (1912–1954): Turing was an inventive logician and mathematician. He conceived of the Turing machine, which is still the major formal model for computation, was crucial to breaking Nazi codes during World War II, and played a major role in the early development of computers. Turing introduced the Turing test as a challenge for artificial intelligence.

Jacques de Vaucanson (1709–1782): A master of automata, Vaucanson created both a flute player and a famous mechanical duck.

Voltaire (1694–1778): Voltaire was a renowned essayist and satirist of the French Enlightenment.

Alfred North Whitehead (1861–1947): Whitehead and Russell's *Principia Mathematica* showed that all of mathematics could be built from a few simple logical concepts and, thereby, paved the road for contemporary computing.

Ludwig Wittgenstein (1889–1951): Wittgenstein was one of the most influential philosophers of the 20th century. Believing that many philosophical problems have their roots in misunderstandings of language, Wittgenstein explored the logic of language with the hope of "dissolving" philosophical problems. His work falls into two distinct periods, marked by the *Tractatus Logico-Philosophicus* and *Philosophical Investigations*. It is the latter work that evokes Analytical Behaviorism, including the private language argument and the parable of the beetles in the boxes.

Bibliography

General Sources and Anthologies:

Beakley, Brian, and Peter Ludlow, eds. *The Philosophy of Mind: Classical Problems/Contemporary Issues*. Cambridge, MA: MIT Press, 1991 (1st ed.), 2006 (2nd ed.). An exhaustive anthology of contemporary work and historical pieces, leaning toward cognitive science.

Block, Ned, Owen Flanagan, and Güven Güzeldere, eds. *The Nature of Consciousness: Philosophical Debates*. Cambridge, MA: MIT Press, 1999. A large but canonical anthology of pieces on consciousness.

Guttenplan, Samuel, ed. *A Companion to the Philosophy of Mind*. Oxford: Blackwell, 1995. Arranged alphabetically by topic, with a helpful introductory mapping of the territory in terms of three levels of analysis.

Heil, John, ed. *Philosophy of Mind: A Guide and Anthology*. Oxford: Oxford University Press, 2004. An anthology of important pieces, historical and contemporary; notable for brief and helpful introductions to each section.

Hofstader, Douglas, and Daniel C. Dennett, eds. *The Mind's I: Fantasies and Reflections on Self and Soul*. New York: Bantam Books, 1981. A rich and enjoyable collection of literature and philosophy, with important pieces mentioned in the lectures by John Searle, Alan M. Turing, Thomas Nagel, and Daniel Dennett.

Velmans, Max, and Susan Schneider, eds. *The Blackwell Companion to Consciousness*. Oxford: Blackwell, 2007. An encyclopedic approach to major theories, authors, and debates.

Interviews:

Baumgartner, Peter, and Sabine Payr, eds. *Speaking Minds: Interviews with Twenty Eminent Cognitive Scientists*. Princeton: Princeton University Press, 1995. Revealing interviews with many of the contemporary figures mentioned in the lectures.

Blackmore, Susan. *Conversations on Consciousness*. New York: Oxford University Press, 2006. A series of informal interviews with many of the contemporary figures mentioned in the lectures.

Works Cited in the Lectures:

Aristotle. *De Interpretatione*, *Prior Analytics*, and *Posterior Analytics*. In *The Basic Works of Aristotle*, edited by Richard McKeon. New York: Random House, 1971. The birth of logic.

Braitenberg, Valentino. *Vehicles: Experiments in Synthetic Psychology*. Cambridge, MA: MIT Press, 1996. A set of short but thought-provoking chapters on how to design simple machines that seem to show complicated psychological behavior.

Brooks, Rodney. "Intelligence without Representation." *Artificial Intelligence* 47 (1991): 139–159. A manifesto of the Brooks approach to artificial intelligence, robotics, and a mind in the world.

Carter, Rita. *Mapping the Mind*. Christopher Frith, scientific adviser. Berkeley: University of California Press, 1999. An absorbing coffee-table book on the brain.

Chalmers, David. "Facing Up to the Hard Problem of Consciousness." *Journal of Consciousness Studies* 2 (1995): 200–219. Reprinted in Heil, *Philosophy of Mind: A Guide and Anthology*, pp. 617–640. The classic piece on the hard problem.

Churchland, Patricia Smith. "Reduction and Antireductionism in Functionalist Theories of Mind." In *Neurophilosophy*. Cambridge, MA: MIT Press, 1986.

Reprinted in Beakley and Ludlow, *Philosophy of Mind*, both editions. In the process of arguing for the philosophical importance of work in neuroscience, Churchland lays out a clear defense of Reductionism.

Churchland, Paul. "Eliminative Materialism and the Propositional Attitudes." *Journal of Philosophy* 78 (1981): 67–90. Reprinted in Heil, *Philosophy of Mind: A Guide and Anthology*, and in Beakley and Ludlow, *Philosophy of Mind*, 2nd ed. A classic statement of the Churchlands' Eliminative Materialism.

Churchland, Patricia Smith, and Paul Churchland. "Could a Machine Think?" *Scientific American* 262 (1990): 32–39. A response piece to Searle's Chinese room thought experiment in the same issue.

Clark, Andy, and David Chalmers, "The Extended Mind." *Analysis* 58 (1998): 10–23. Reprinted in *The Philosopher's Annual*, vol. 21, edited by Patrick Grim, Ken Baynes, and Gary Mar. Atascadero, CA: Ridgeview Press, 1999. Clark and Chalmers's elegant statement of the "extended mind" thesis.

Crevier, Daniel. *AI: The Tumultuous History of the Search for Artificial Intelligence*. New York: Bantam Books, 1993. An entertaining and informative treatment of debates that continue within artificial intelligence.

Crick, Francis. *The Astonishing Hypothesis: The Scientific Search for the Soul*. New York: Touchstone, 1994. An accessible source for a number of topics in neuroscience, including Crick and Koch's theory of consciousness as synchronized 40-Hertz firing across the brain.

Damasio, Antonio. *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: Penguin Books, 1994. Contemporary neuroscience applied to questions in philosophy and psychology; as engaging as it is informative.

Dennett, Daniel. *Brainstorms*. Cambridge, MA: MIT Press, 1981. A collection of Dennett's earlier articles in philosophy of mind, ending with the wonderful "Where Am I?" Accessible and thought-provoking.

———. *Elbow Room: The Varieties of Free Will Worth Wanting*. Cambridge, MA: MIT Press, 1984. Although Dennett has more recent work on free will, this is his classic treatment and, in many ways, still the best.

———. *Freedom Evolves*. New York: Penguin Books, 2003. As always, Dennett is eloquent and entertaining. Here, he outlines Compatibilism in an evolutionary setting.

———. "The Nature of Images and the Introspective Trap." In *Content and Consciousness*. London: Routledge & Kegan Paul; New York: Humanities Press, 1969. Reprinted in Beakley and Ludlow, *Philosophy of Mind*, 1st ed. A classic Dennett attack on the inner theater.

———. "Quining Qualia." In *Consciousness in Modern Science*, edited by A. Marcel and E. Bisiach. Oxford: Oxford University Press, 1988. Reprinted in Block, Flanagan, and Güzeldere, *The Nature of Consciousness: Philosophical Debates*. A difficult but rewarding piece to read repeatedly.

Dennett, Daniel, and Marcel Kinsbourne. "Time and the Observer: The Where and When of Consciousness in the Brain." *Behavioral and Brain Sciences* 15 (1992): 183–247. Reprinted in *The Philosopher's Annual*, vol. 15, edited by Patrick Grim, Gary Mar, and Peter Williams. Atascadero, CA: Ridgeview Press, 1994. Also reprinted in Block, Flanagan, and Güzeldere, *The Nature of Consciousness: Philosophical Debates.* The full development of several objections to the inner theater.

Descartes, René. *Meditations on First Philosophy*. John Cottingham, ed. Cambridge: Cambridge University Press, 1996. Available in many editions, excerpted in many anthologies. A primary source for Cartesian Dualism.

Durrell, Lawrence. *The Alexandria Quartet: Justine*, *Balthazar*, *Mountolive*, and *Clea*. London: Faber and Faber, 1962. This set of beautiful novels makes the point that no behavioral description can be interpretationally exhaustive.

Fodor, Jerry. "Observation Reconsidered." *Philosophy of Science* 51 (1984): 23–43. Reprinted in *The Philosopher's Annual*, vol. 7, edited by Patrick Grim, Christopher J. Martin, and Michael A. Simon. Atascadero, CA:

Ridgeview Press, 1984. An effective response to overstatements of holism and the influence of belief on perception.

Gardner, Howard. *Frames of Mind: The Theory of Multiple Intelligences*. New York: Basic Books, 1983. A complete but readable outline of his theory, with progressive chapters on different intelligences.

———. *Multiple Intelligences: The Theory in Practice*. New York: Basic Books, 1993. An accessible introduction to his theory and its implications.

Gibson, James J. "Autobiography." In *Reasons for Realism: Selected Essays of James J. Gibson*, edited by Edward Reed and Rebecca Jones, pp. 7–22. London: Lawrence Erlbaum, 1982. A brief but effective introduction to major elements of Gibson's approach.

———. *The Ecological Approach to Visual Perception*. London: Lawrence Erlbaum, 1986. Gibson's major work in applying the concept of perceptual affordances.

Gibson, William, and Bruce Sterling, *The Difference Engine*. New York: Bantam Books, 1991. A novelistic look at what history might have been like if Charles Babbage had succeeded in building steam-driven computers in 1840.

Gould, Stephen Jay. *The Mismeasure of Man*. New York: W.W. Norton, 1981. Gould's outspoken criticism of IQ testing.

Grim, Patrick. "Free Will in Context: A Contemporary Philosophical Perspective." *Behavioral Sciences and the Law* 25 (2007): 183–201. A survey of current approaches with a concluding outline of contextualism.

Hanson, N. R. "Observation." In *Patterns of Discovery*. Cambridge: Cambridge University Press, 1958. Excerpted in *Introductory Readings in Philosophy of Science*, edited by E. D. Klemke, Robert Hollinger, and A. David Kline. Amherst, NY: Prometheus Books, 1980. A strong classic statement regarding theory-laden perception.

Hauser, Marc D. *Wild Minds: What Animals Really Think*. New York: Henry Holt and Co., 2000. An entertaining discussion of animal cognition that introduces a range of important contemporary research.

Hoffman, Donald D. *Visual Intelligence: How We Create What We See*. New York: W.W. Norton & Company, 1998. A fascinating exploration of optical illusions and what they tell us about perception.

Humphrey, Nicholas. *Seeing Red: A Study in Consciousness*. Cambridge, MA: Harvard University Press, 2006. A short and popular introduction to issues in perception by the man who first studied blindsight in monkeys.

Jackson, Frank. "Epiphenomenal Qualia." *Philosophical Quarterly* 32 (1982): 127–136. The original appearance of black-and-white Mary. A good general discussion of the argument appears in Daniel Stoljar and Yujin Nagasawa's introduction to *There's Something about Mary*.

Kuhn, Thomas S. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press, 1962; with postscript, 1970. An extremely influential piece of work in philosophy of science.

Kurzweil, Raymond. *The Singularity Is Near: When Humans Transcend Biology*. New York: Penguin, 2005. Kurzweil's predictions regarding the future of artificial intelligence.

Locke, John. *An Essay Concerning Human Understanding*. New York: Dover Publications, 1959. A historical classic of Empiricism, as difficult to read as most things published in 1689.

Loftus, Elizabeth F. *Eyewitness Testimony*. Cambridge, MA: Harvard University Press, 1979. The classic text regarding the fallibility of eyewitness testimony.

McGinn, Colin. "Can We Solve the Mind-Body Problem?" *Mind* 98 (1989): 349–366. Reprinted in Block, Flanagan, and Güzeldere, *The Nature of Consciousness: Philosophical Debates*. A clearly posed problem with some obscurity in McGinn's answer to it.

Miedaner, Terrell. “The Soul of Martha, a Beast” and “The Soul of the Mark III Beast.” In *The Soul of Anna Klane*. New York: Ballantine Books, 1978. Reprinted in Hofstadter and Dennett, *The Mind’s I*. Taken together, a wonderful pair of fictional reflections on life and consciousness—human, animal, and mechanical.

Minsky, Marvin, and Seymour Papert. *Perceptrons, Expanded Edition*. Cambridge, MA: MIT Press, 1988. In its original form, this is the book that interrupted fruitful work on neural networks for a generation.

Moravec, Hans. *Mind Children: The Future of Human and Robot Intelligence*. Cambridge, MA: Harvard University Press, 1988. Moravec envisages a beautiful future in which our machines replace us.

Nagel, Thomas. *Mortal Questions*. New York: Cambridge University Press, 1979. A collection of Nagel’s essays, with emphasis on ethics. Nagel is always clear, direct, and surprising.

———. *The View from Nowhere*. Oxford: Oxford University Press, 1986. Nagel’s defining statement on issues of subjectivity and objectivity across the philosophical spectrum.

———. “What Is It Like to Be a Bat?” *The Philosophical Review* 83 (1974): 435–350. Reprinted in Nagel, *Mortal Questions*, and in Hofstadter and Dennett, *The Mind’s I*. Nagel’s seminal article on the problem of consciousness.

Noë, Alva. *Action in Perception*. Cambridge, MA: MIT Press, 2004. Noë’s enactive approach is a philosophical extension of the Gibsonian tradition.

Parfit, Derek. *Reasons and Persons*. Oxford: Clarendon Press, 1984. Even better than its sterling reputation would lead one to expect. Densely argued, Parfit’s work calls for reading in small bits and thinking at large.

Penrose, Roger. *The Emperor’s New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford: Oxford University Press, 1989. The first of

Penrose's two books on undecidability and the limits of algorithms, quantum mechanics, and the brain. Dense but often rewarding.

———. *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford: Oxford University Press, 1994. The second of Penrose's two books on undecidability and the limits of algorithms, quantum mechanics, and the brain. Dense but often rewarding.

Place, U. T. "Is Consciousness a Brain Process?" *British Journal of Psychology* 47 (1856): 44–50. Reprinted in Beakley and Ludlow, *Philosophy of Mind*, both editions. A classic statement of the claim that mental states are brain states.

Putnam, Hilary. "The Nature of Mental States." Originally published as "Psychological Predicates" in *Art, Mind, and Religion*, edited by W. Capitan and D. D. Merrill. Pittsburgh: University of Pittsburgh Press, 1967. Reprinted in Beakley and Ludlow, *Philosophy of Mind*, both editions, and as "Psychological Predicates" in Heil, *Philosophy of Mind: A Guide and Anthology*. A classic statement of Functionalism.

Ramachandran, V. S., and Sandra Blakeslee. *Phantoms in the Brain: Probing the Mysteries of the Human Mind*. New York: William Morrow, 1988. An engaging discussion of neuroscience and its lessons regarding the mind, with a focus on Ramachandran's research regarding phantom limbs.

Rosenthal, David. "A Theory of Consciousness." In Block, Flanagan, and Güzeldere, *The Nature of Consciousness: Philosophical Debates*. A brief and effective presentation of Rosenthal's higher-order thought (HOT) theory of consciousness.

Rumelhart, David E., James L. McClelland, and the PDP Research Group. *Parallel Distributed Processing*, vol. 1: *Foundations*. Cambridge, MA: MIT Press, 1988. The resurrection of neural nets. Only for the initiated.

Russell, Bertrand. *Philosophy*. New York: W.W. Norton, 1927. One of Russell's many works; a bit dated but wonderfully written and always thought-provoking.

———, and Alfred North Whitehead. *Principia Mathematica*. Cambridge: Cambridge University Press, 1927. First published in 1910–1913, this is Russell and Whitehead's demonstration that all of mathematics could be generated from a few logical symbols and rules. Even for the specialist, now of purely historical interest.

Ryle, Gilbert. *The Concept of Mind*. New York: Barnes and Noble, 1950; Chicago: University of Chicago Press, 2002 (a new edition with an introduction by Daniel C. Dennett). Excerpted in Beakley and Ludlow, *Philosophy of Mind*, both editions. The classical statement of Analytical Behaviorism.

Sacks, Oliver. *An Anthropologist on Mars: Seven Paradoxical Tales*. New York: Vintage Books, 1996. Entertaining and enlightening.

———. *The Man Who Mistook His Wife for a Hat*. New York: HarperCollins, 1985; New York: Touchstone 1998. A thoroughly fascinating set of cases, with Sacks's speculations as to what they tell us about the mind.

Searle, John. "Is the Brain's Mind a Computer Program?" *Scientific American* 262 (1990): 26–31. The Chinese room argument in *Scientific American* form, accompanied by a response from the Churchlands.

———. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (1980): 417–450. Although Searle's text is reprinted in several of the resources listed here (Beakley and Ludlow, *Philosophy of Mind*, both editions; Heil, *Philosophy of Mind: A Guide and Anthology*; and Hofstadter and Dennett, *The Mind's I*), it is best read in its original context, with extensive comments by others and his replies.

Stoljar, Daniel, and Yujin Nagasawa. "Introduction." In *There's Something about Mary*, edited by Peter Ludlow, Yujin Nagasawa, and Daniel Stoljar. Cambridge, MA: MIT Press, 2004. A good overview of Frank Jackson's black-and-white Mary argument.

Vacca, John, ed. *The World's 20 Greatest Unsolved Problems*. Upper Saddle River, NJ: Prentice-Hall, 2005. One of the few books that focuses on what we *don't* know.

Wittgenstein, Ludwig. *The Blue and Brown Books*. New York: Harper and Row; Oxford: Blackwell, 1958. Compiled from students' notes originally distributed in blue and brown covers in mimeograph form, *The Blue Book* offers a good introduction to the private language argument in particular and Wittgenstein's intriguing obscurity in general.

———. *Philosophical Investigations*. New York: Macmillan, 1958. The later Wittgenstein in all his aphoristic obscurity, simultaneously intriguing and infuriating.

Wood, Gaby. *Edison's Eve: A Magical History of the Quest for Mechanical Life.* New York: Anchor Books, 2002. Rich initial chapters on the legend of Descartes' robotic daughter and Vaucanson's automata; somewhat sketchier later on.

Additional Resources:

Baars, Bernard J., William P. Banks, and James B. Newman, eds. *Essential Sources in the Scientific Study of Consciousness.* Cambridge, MA: MIT Press, 2003. Purely for the specialist.

Boden, Margaret A. *Mind as Machine: A History of Cognitive Science*, vols. 1 and 2. Oxford: Clarendon Press, 2006. A vast, detailed, opinionated but valuable research source.

Uribe, Diego. *Truly Baffling Optical Illusions*. New York: Sterling Publishing, 2003. One of the best of many collections of optical illusions, with some discussion of why certain effects occur.

Williams, Michael R. *A History of Computing Technology*. Los Alamitos, CA: IEEE Computer Science Press, 1997. An exhaustive compendium on the history of calculational notations, techniques, and machines. One of the best.

Film and Video Sources:

Lang, Fritz. *Metropolis*. 1927.

Scott, Ridley. *Blade Runner*. 1982.

The following are classic reflections on minds, machines, and society:

Morris, Erroll. *Fast, Cheap, and Out of Control.* 1997. A cult documentary that mixes footage and interviews regarding lion taming, topiary, naked mole rats, and Rodney Brooks on robots.

Rose, Reginald, and Sidney Lumet. *Twelve Angry Men*. 1957. A psychological drama that makes memorable points regarding prejudice and misperception. Make sure you see the black-and-white version starring Henry Fonda.

Scientific American Frontiers. *Robots Alive!*, *Robot Pals*, *Changing Your Mind*. 1997. Alan Alda's entertaining programs on topics covered in the lectures.

Internet Sources:

Bach, Michael. *77 Optical Illusions and Visual Phenomena*. http://www.michaelbach.de/ot/.

Bongard, Josh, Victor Zykov, and Hod Lipson. *Resilient Machines through Continuous Self-Modeling*. http://ccsl.mae.cornell.edu/ research/selfmodels/. The star robot in action.

Chalmers, David. This professor and author offers an astounding set of Web pages regarding consciousness and philosophy of mind. When does this man have time to do anything else? Chalmers's sites include the following:

Online Papers on Consciousness. http://consc.net/online.html.

Mindpapers: A Bibliography of the Philosophy of Mind and the Science of Consciousness. http://consc.net/biblio.html. Extensive.

Zombies on the Web. http://consc.net/zombies.html. A compilation of sites on zombies.

Computer History Museum. http://www.computerhistory.org/ about_us.html.

Deutsch, Diana. http://psy.ucsd.edu/~ddeutsch/. Auditory illusions.

Duke University, Neurobiology, Laboratory of Dale Purves, M.D. *Color Contrast: Cube.* http://www.neuro.duke.edu/faculty/purves/gallery9.html. Shows the color contrast experiment referred to in Lecture Twenty. Astoundingly, the blue squares on the top surface in the left image are, in fact, identical to the yellow squares on the top surface in the right image. Taken out of context, both are an identical shade of gray. The experiment also appears with some additional discussion at http://discovermagazine.com/2004/feb/ neuroquest/.

e-Chalk Optical Illusions. http://www.echalk.co.uk/amusements/Optical Illusions/illusions.htm. Includes interactive forms of the color contrast experiment and other color illusions.

Exploratorium: The Museum of Science, Art and Human Perception. http://www.exploratorium.edu/seeing/exhibits/index.html. Includes a number of visual illusions.

Home Page of The Loebner Prize in Artificial Intelligence. http://www.loebner.net/Prizef/loebner-prize.html. Includes information on the Loebner Prize and links to recent winning programs that you can review.

Macmillan, Malcolm. *The Phineas Gage Information Page*. http://www.deakin.edu.au/hmnbs/psychology/gagepage/.

Science Museum, http://www.sciencemuseum.org.uk/visitmuseum/galleries/computing/ondisplay.aspx.

Shore, David I. *Measuring Auditory Saltation*. http://www.mohsho.com/dshore/sound_top.html. A version of the auditory rabbit.

T.I.L. Productions. *Vaucanson and His Remarkable Automatons*. http://www.automates-anciens.com/english_version/automatons-music-boxes/vaucanson-automatons-androids.php. Site shows what may be photos of the remains of Vaucanson's duck.

University of North Carolina at Charlotte. http://www.philosophy.uncc.edu/faculty/phi/Phi_Color2.html. An interactive exploration of the color phenomenon.

Visual Cognition Lab, University of Illinois. http:// viscog.beckman.uiuc.edu/djs_lab/demos.html. Includes the unseen gorilla video mentioned in Lecture Nineteen, together with wonderful exercises regarding change blindness.

Zalta, Edward N., principal ed. *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/. An unequaled website source on all philosophical topics.

Permissions Acknowledgments