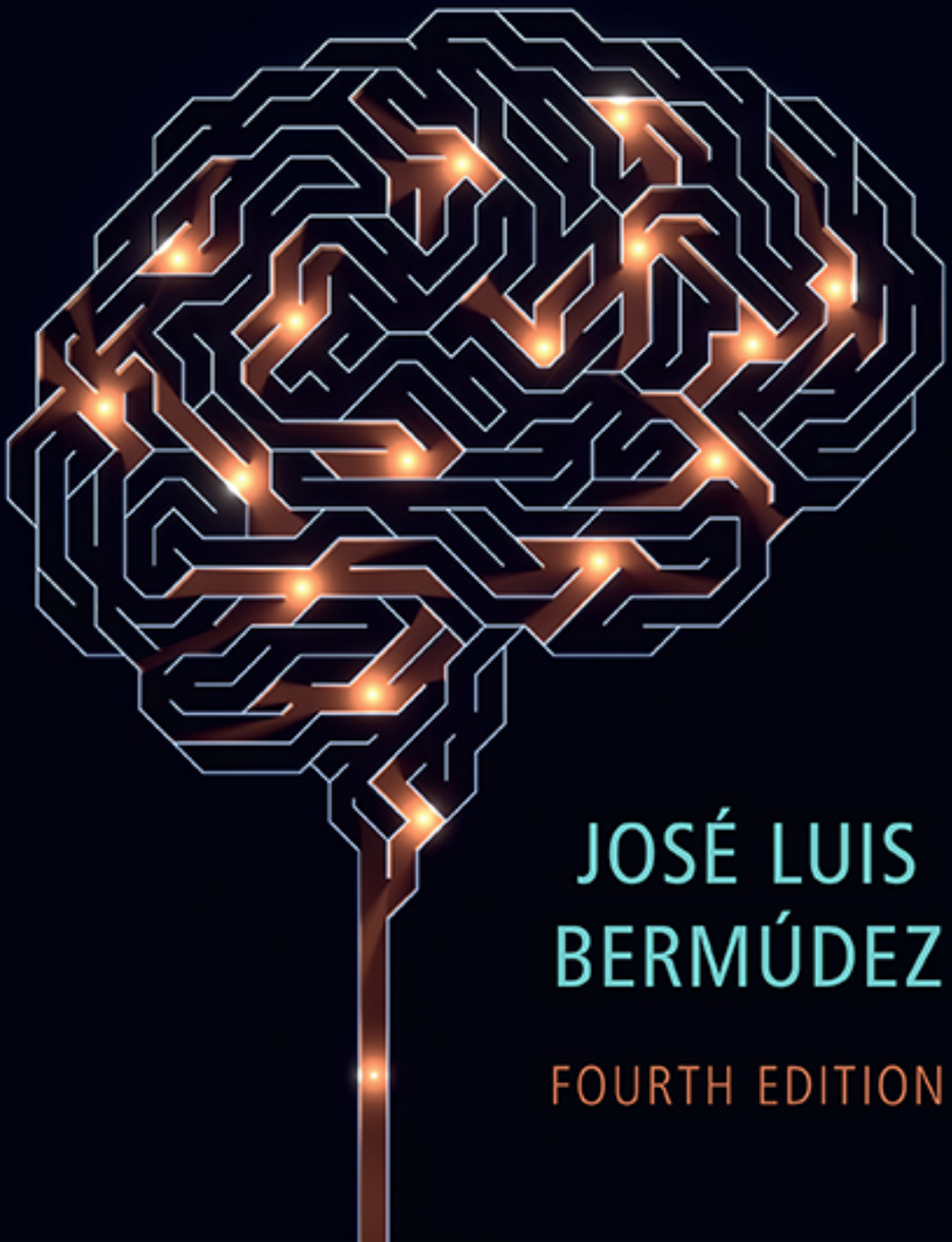
An Introduction to the Science of the Mind

Cognitive Science



JOSÉ LUIS BERMÚDEZ

FOURTH EDITION

Copyright statement
Cognitive Science

By José Luis Bermúdez

ISBN 13: 9781009081054

Publisher: Cambridge University Press (2014)

This copy is made under the terms of the Copyright and Rights in Performances (Disability) Regulations 2014 to be used only by a print-disabled person. Except as permitted by law, it may not be further copied, nor may it be supplied to any other person, without permission.

C O G N I T I V E S C I E N C E

Fourth Edition

The fourth edition of this popular text has been significantly rewritten to make it more accessible to students and easier for instructors to use. It remains distinctive in presenting a unified narrative of cognitive science as a field of inquiry in its own right. Thematically organized, *Cognitive Science* underscores the problems and solutions of cognitive science as an interdisciplinary enterprise rather than more narrowly examining individually the subjects that contribute to it — psychology, neuroscience, linguistics, and so on. The generous use of examples, illustrations, and applications demonstrates how theory and experiment can be applied to unlock the mysteries of the human mind. Drawing upon cutting-edge research, the text has been updated and enhanced with a new chapter on emotions and the emerging field of affective science. An extensive online set of resources is available to aid both instructors and students.

José Luis Bermúdez′s research interests are interdisciplinary in nature at the intersection of philosophy, psychology, and neuroscience. His many books include *The Paradox of Self-Consciousness* (MIT Press, 1998), which analyzes the nature of self-awareness; *Thinking without Words* (Oxford University Press, 2003), which offers a model for thinking about the cognitive achievements and abilities of prelinguistic infants and nonlinguistic humans; and *Decision Theory and Rationality* (Oxford University Press, 2009), which explores tensions in how the concept of rationality is defined and formalized in different academic disciplines. Recent books include *Understanding "I": Language and Thought* (Oxford University Press, 2017) and *The Bodily Self: Selected Essays* (MIT Press, 2018). His most recent book, *The Power of Frames* (Cambridge University Press, 2020), was supported by an American Council of Learned Societies (ACLS) fellowship for 2018–2019 and a National Endowment for the Humanities (NEH) Summer Stipend for 2018. A target article with peer commentary based on *The Power of Frames* appeared in *Behavioral and Brain Sciences* (2022).

COGNITIVE SCIENCE

An Introduction to the Science of the Mind

Fourth Edition

José Luis Bermúdez
Texas A&M University



CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314 – 321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05 – 06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of the University of Cambridge.

It furthers the University ' s mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/highereducation/isbn/9781316513378

DOI: 10.1017/9781009064880

First and second editions © José Luis Bermúdez 2010, 2014

Third edition © Cambridge University Press 2020

Fourth edition © José Luis Bermúdez 2023

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published 2010

Second edition published 2014

Third edition published 2020

3rd printing 2021

Fourth edition published 2023

Printed in the United Kingdom by TJ Books Limited, Padstow, Cornwall, 2023

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

ISBN 978-1-316-51337-8 Hardback

ISBN 978-1-009-07367-7 Paperback

Additional resources for this publication at www.cambridge.org/bermudez4e

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

Brief Contents

List of Figures
List of Tables
List of Boxes
Preface
Acknowledgments

Introduction: The Challenge of Cognitive Science

Part I HISTORICAL LANDMARKS

1 The Prehistory of Cognitive Science

2 The Discipline Matures: Three Milestones

3 The Turn to the Brain

Part II MODELS AND TOOLS

4 Physical Symbol Systems and the Language of Thought

5 Neural Networks and Distributed Information Processing

6 Applying Dynamical Systems Theory to Model the Mind

7 Bayesianism in Cognitive Science

8 Modules and Architectures

9 Strategies for Brain Mapping

Part III APPLICATIONS

10 Models of Language Learning

11 Object Perception and Folk Physics

12 Machine Learning: From Expert Systems to Deep Learning

13 Exploring Mindreading

14 Robotics: From GOFAI to Situated Cognition and Behavior-Based Robotics

15 The Cognitive Science of Consciousness

16 The Emotions: From Cognitive Science to Affective Science

17 Looking Ahead: Challenges and Opportunities

Glossary

Bibliography

Index

Contents

List of Figures
List of Tables
List of Boxes
Preface
Acknowledgments

Introduction: The Challenge of Cognitive Science
Overview
Cognitive Science: An Interdisciplinary Endeavor
Levels of Explanation: The Contrast between Psychology and Neuroscience
How Psychology Is Organized
How Neuroscience Is Organized
The Challenge of Cognitive Science
Three Dimensions of Variation
The Space of Cognitive Science

Part I HISTORICAL LANDMARKS

1 The Prehistory of Cognitive Science
Overview
1.1 The Reaction against Behaviorism in Psychology
Learning without Reinforcement: Tolman and Honzik, "' Insight ' in Rats " (1930)

Cognitive Maps in Rats? Tolman et al., " Studies in Spatial Learning " (1946)

Plans and Complex Behaviors: Lashley, " The Problem of Serial Order in Behavior " (1951)

1.2 The Theory of Computation and the Idea of an Algorithm

Algorithms and Turing Machines: Turing, " On Computable Numbers, with an Application to the Decision Problem " (1936 – 1937)

1.3 Linguistics and the Formal Analysis of Language

The Structure of Language: Chomsky ' s *Syntactic Structures* (1957)

1.4 Information-Processing Models in Psychology

How Much Information Can We Handle? George Miller ' s " The Magical Number Seven, Plus or Minus Two " (1956)

The Flow of Information: Donald Broadbent ' s " The Role of Auditory Localization in Attention and Memory Span " (1954) and *Perception and Communication* (1958)

1.5 Connections and Points of Contact

Summary

Checklists

2 The Discipline Matures: Three Milestones

Overview

2.1 Language and Micro-Worlds

Natural Language Processing: Winograd, *Understanding Natural Language* (1972)

SHRDLU in Action

2.2 How Do Mental Images Represent?

Mental Rotation: Shepard and Metzler, " Mental Rotation of Three-Dimensional Objects " (1971)

Information Processing in Mental Imagery

2.3 An Interdisciplinary Model of Vision

Levels of Explanation: Marr ' s *Vision* (1982)

Applying Top-Down Analysis to the Visual System

Summary

Checklists

3 The Turn to the Brain

Overview

3.1 Cognitive Systems as Functional Systems?

3.2 The Anatomy of the Brain and the Primary Visual Pathway

The Two-Visual-Systems Hypothesis: Ungerleider and Mishkin, " Two Cortical Visual Systems " (1982)

3.3 Extending Computational Modeling to the Brain

A New Set of Algorithms: Rumelhart et al., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (1986)

Pattern Recognition in Neural Networks: Gorman and Sejnowski, " Analysis of Hidden Units in a Layered Network Trained to Identify Sonar Targets " (1998)

3.4 Mapping the Stages of Lexical Processing

Functional Neuroimaging with PET

Petersen et al., " Positron Emission Tomographic Studies of the Cortical Anatomy of Single-Word Processing " (1988)

3.5 Studying Memory for Visual Events

Functional Neuroimaging with fMRI

Brewer et al., " Making Memories: Brain Activity That Predicts How Well Visual Experience Will Be Remembered " (1998)

3.6 The Neural Correlates of the BOLD Signal

Logothetis, " The Underpinnings of the BOLD Functional Magnetic Resonance Imaging Signal " (2001)

Summary

Checklists

Part II MODELS AND TOOLS

4 Physical Symbol Systems and the Language of Thought

Overview

4.1 The Physical Symbol System Hypothesis

Symbols and Symbol Systems

Transforming Symbol Structures

Intelligent Action and the Physical Symbol System

4.2 From Physical Symbol Systems to the Language of Thought

Intentional Realism and Causation by Content

The Language of Thought and the Relation between Syntax and Semantics

4.3 The Russian Room Argument and the Turing Test

Responding to the Russian Room Argument

Summary

Checklists

5 Neural Networks and Distributed Information Processing

Overview

5.1 **Neurally Inspired Models of Information Processing**

Neurons and Network Units

5.2 **Single-Layer Networks and Boolean Functions**

Learning in Single-Layer Networks: The Perceptron-Convergence Rule

Linear Separability and the Limits of Perceptron Convergence

5.3 **Training Multilayer Networks**

The Backpropagation Algorithm

How Biologically Plausible Are Neural Networks?

5.4 **Information Processing in Neural Networks: Key Features**

Distributed Representations

No Clear Distinction between Information Storage and Information Processing

The Ability to Learn from " Experience "

Summary

Checklists

6 Applying Dynamical Systems Theory to Model the Mind

Overview

6.1 **Cognitive Science and Dynamical Systems**

What Are Dynamical Systems?

The Dynamical Systems Hypothesis: Cognitive Science without Representations?

6.2 Applying Dynamical Systems: Two Examples from Child Development

Two Ways of Thinking about Motor Control

Dynamical Systems and the A-Not-B Error

Assessing the Dynamical Systems Approach

Summary

Checklists

7 Bayesianism in Cognitive Science

Overview

7.1 Bayesianism: A Primer

Degrees of Belief and Subjective Probability

Conditional Probability

Bayes ' s Rule (the Short Version)

7.2 Perception as a Bayesian Problem

The Predictive Challenge of Perception

Case Study: Binocular Rivalry

7.3 Neuroeconomics: Bayes in the Brain

What Is Expected Utility?

Case Study: Neurons That Code for Expected Utility

Summary

Checklists

8 Modules and Architectures

Overview

8.1 Architectures for Artificial Agents

Three Agent Architectures

8.2 **Fodor on the Modularity of Mind**
Modular and Nonmodular Processing
8.3 **The Massive Modularity Hypothesis**
The Cheater-Detection Module
The Evolution of Cooperation
Two Arguments
Evaluating the Arguments for Massive Modularity
8.4 **Hybrid Architectures: The Example of ACT-R**
The ACT-R Architecture
ACT-R as a Hybrid Architecture
Summary
Checklists

9 Strategies for Brain Mapping
Overview
9.1 **Structure and Function in the Brain**
Exploring Anatomical Connectivity
9.2 **Studying Cognitive Functioning: Techniques from Neuroscience**
Mapping the Brain ' s Electrical Activity: EEG and MEG
Mapping the Brain ' s Blood Flow and Blood Oxygen Levels: PET and fMRI
9.3 **Combining Resources I: The Locus-of-Selection Problem**
Combining ERPs and Single-Unit Recordings
9.4 **Combining Resources II: Networks for Attention**
Two Hypotheses about Visuospatial Attention
9.5 **From Data to Maps: Problems and Pitfalls**

From Blood Flow to Cognition?
Noise in the System?
Functional Connectivity versus Effective Connectivity
Summary
Checklists

Part III APPLICATIONS

10 Models of Language Learning
Overview
10.1 Language and Rules
Understanding a Language and Learning a Language
10.2 Language Learning and the Language of Thought: Fodor's Argument
10.3 Language Learning in Neural Networks
The Challenge of Tense Learning
Connectionist Models of Tense Learning
10.4 Bayesian Language Learning
Probabilities in Word and Phrase Segmentation
Understanding Pronouns
Learning Linguistic Categories
Summary
Checklists

11 Object Perception and Folk Physics
Overview

11.1 Object Permanence and Physical Reasoning in Infancy

Infant Cognition and the Dishabituation Paradigm

How Should the Dishabituation Experiments Be Interpreted?

11.2 Neural Network Models of Children's Physical Reasoning

Modeling Object Permanence

Modeling the Balance-Beam Problem

11.3 Conclusion: The Question of Levels

Summary

Checklists

12 Machine Learning: From Expert Systems to Deep Learning

Overview

12.1 Expert Systems and Machine Learning

Expert Systems and Decision Trees

ID3: An Algorithm for Machine Learning

12.2 Representation Learning and Deep Learning

Deep Learning and the Visual Cortex

12.3 The Machinery of Deep Learning

Autoencoders

Convolutional Neural Networks

12.4 Deep-Reinforcement Learning

Summary

Checklists

13 Exploring Mindreading

Overview

13.1 Pretend Play and Metarepresentation

Leslie on Pretend Play and Metarepresentation

13.2 Metarepresentation, Autism, and Theory of Mind

Using the False-Belief Task to Study Mindreading

Interpreting the Results

Implicit and Explicit Understanding of False Belief

13.3 The Mindreading System

First Steps in Mindreading

From Dyadic to Triadic Interactions: Joint Visual Attention

TESS and TOMM

13.4 Mindreading as Simulation

Standard Simulationism

Radical Simulationism

13.5 The Cognitive Neuroscience of Mindreading

Neuroimaging Evidence for a Dedicated Theory-of-Mind System?

Neuroscientific Evidence for Simulation in Low-Level Mindreading?

Neuroscientific Evidence for Simulation in High-Level Mindreading?

Summary

Checklists

14 Robotics: From GOFAI to Situated Cognition and Behavior-Based Robotics

Overview

14.1 GOFAI Robotics: SHAKEY

SHAKEY ' s Software I: Low-Level Activities and Intermediate-Level Actions

SHAKEY ' s Software II: Logic Programming in STRIPS and PLANEX

14.2 Situated Cognition and Biorobotics

The Challenge of Building a Situated Agent

Situated Cognition and Knowledge Representation

Biorobotics: Insects and Morphological Computation

14.3 From Subsumption Architectures to Behavior-Based Robotics

Subsumption Architectures: The Example of Allen

Behavior-Based Robotics: TOTO

Multiagent Programming: The Nerd Herd

Summary

Checklists

15 The Cognitive Science of Consciousness

Overview

15.1 Information Processing without Conscious Awareness: Some Basic Data

Consciousness and Priming

Nonconscious Processing in Blindsight and Unilateral Spatial Neglect

15.2 So What Is Consciousness For?

What Is Missing in Blindsight and Spatial Neglect

Milner and Goodale: Vision for Action and Vision for Perception

What Is Missing in Masked Priming

15.3 Two Types of Consciousness and the Hard Problem

15.4 Theories of Consciousness

Higher-Order Theories

The Global Workspace Theory

The Global Neuronal Workspace Theory

15.5 Conclusion

Summary

Checklists

16 The Emotions: From Cognitive Science to Affective Science

Coauthored with Dong An

Overview

16.1 Early Theories

Physical Symbol Systems and the Emotions

Ekman and Basic Emotions

16.2 Affective Space and the Affective Scientist's Toolkit

Elements of an Emotional Episode

Genetic Tools in Affective Science

Lesion Studies

16.3 Fear: A Multilevel and Multidisciplinary Case Study

Fear Conditioning and Fear Responses in Rodents

Fear and Amygdala Damage in Humans

Neuroimaging of Fear in Humans

Summary

Checklists

17 Looking Ahead: Challenges and Opportunities

Overview

17.1 Exploring the Connectivity of the Brain: The Human Connectome Project and Beyond

17.2 Understanding What the Brain Is Doing When It Appears Not to Be Doing Anything

17.3 Neural Prosthetics

17.4 Cognitive Science and the Law

17.5 Autonomous Vehicles: Combining Deep Learning and Intuitive Knowledge

Glossary

Bibliography

Index

Figures

0.1 Connections among the cognitive sciences, as depicted in the Sloan Foundation ' s 1978 report

0.2 Some of the principal branches of scientific psychology

0.3 Levels of organization and levels of explanation in the nervous system

0.4 The spatial and temporal resolution of different tools and techniques in neuroscience

0.5 The " space " of contemporary cognitive science

1.1 A rat in a Skinner box

1.2 A fourteen-unit T-Alley maze (measurements in inches)

1.3 A cross-maze, as used by Tolman et al. (1946)

1.4 Schematic representation of a Turing machine

1.5 A sample phrase-structure tree for the sentence " John has hit the ball. "

1.6 Donald Broadbent ' s (1958) model of selective attention

2.1 A question for SHRDLU about its virtual micro-world

2.2 An algorithm for determining whether a given input is a sentence or not

2.3 Algorithms for identifying noun phrases and verb phrases

2.4 Procedure for applying the command CLEARTOP

2.5 SHRDLU acting on the initial command to pick up a big red block

2.6 SHRDLU completing instruction 3 in the dialog: “ Find a block that is taller than the one you are holding and put it in the box. ”

2.7 Examples of the three-dimensional figures used in Shepard and Metzler ' s (1971) studies of mental rotation

2.8 Results of Shepard and Metzler ' s (1971) studies of mental rotation

2.9 Examples of vertically and horizontally oriented objects that subjects were asked to visualize in Kosslyn ' s (1973) scanning study

2.10 Two images of a bucket

2.11 Two examples of Marr ' s primal sketch, the first computational stage in his analysis of the early visual system

2.12 An example of part of the 2.5D sketch

2.13 An illustration of Marr ' s 3D sketch, showing how the individual components are constructed

2.14 The place of the implementational level within Marr ' s overall theory

2.15 An illustration of the hierarchical organization of the visual system, including which parts of the brain are likely responsible for processing different types of visual information

3.1 The large-scale anatomy of the brain, showing the forebrain,

the midbrain, and the hindbrain

3.2 A vertical slice of the human brain, showing the cerebrum

3.3 The division of the left cerebral hemisphere into lobes

3.4 The primary visual pathway

3.5 Image showing ventral (*orange*) and dorsal (*red*) pathways in the human visual system

3.6 Design and results of Ungerleider and Mishkin ′ s cross-lesion disconnection studies

3.7 A generic three-layer connectionist network (also known as an *artificial neural network*)

3.8 Gorman and Sejnowski ′ s mine-/rock-detector network

3.9 Images showing the different areas of activation (as measured by blood flow) during the four different stages in Petersen et al. ′ s (1988) lexical access studies

3.10 A flowchart relating areas of activation in Petersen et al. ′ s (1988) study to different levels of lexical processing

3.11 Neural areas showing activity when subjects looked at pictures

3.12 Neural areas where activation is correlated with levels of memory performance

3.13 A microelectrode making an extracellular recording

3.14 Simultaneous microelectrode and fMRI recordings from a cortical site showing the neural response to a pulse stimulus of twenty-four seconds

4.1 A typical traveling salesperson problem

4.2 The structure of Fodor ' s argument for the language-of-thought hypothesis

4.3 Inside and outside the Russian room

5.1 Schematic illustration of a typical neuron

5.2 An artificial neuron

5.3 Four different activation functions

5.4 Illustration of a mapping function

5.5 A single-layer network representing the Boolean function AND

5.6 A single-layer network representing the Boolean function NOT

5.7 The starting configuration for a single-layer network being trained to function as a NOT gate through the perceptron-convergence rule

5.8 Graphical representations of the AND and XOR (exclusive-OR) functions, showing the linear separability of AND

5.9 A multilayer network representing the XOR (exclusive-OR) function

5.10 The computational operation performed by a unit in a connectionist model

6.1 The trajectory through the state space of a swinging pendulum

6.2 The state space of a swinging pendulum in a three-dimensional state space

6.3 Illustration of the Watt governor, together with a schematic representation of how it works

6.4 An example of the computational approach to motor control

6.5 The stage IV search task, which typically gives rise to the A-not-B error in infants at around the age of nine months

6.6 An infant sitting for an A trial (*left*) and standing for a B trial (*right*)

6.7 Applying the dynamic field model to the A-not-B error

7.1 An illustration purporting to be of Thomas Bayes from a 1936 book on the history of life insurance

7.2 A diagram showing (a) the proportion of the probability space in which A is true, (b) the proportion of the probability space in which B is true, and (c) the intersection of A and B (which is the region where A and B are both true)

7.3 Four of the seven Gestalt principles of grouping, illustrated and explained

7.4 Two examples of stimuli used to elicit binocular rivalry

7.5 Two well-known ambiguous figures: (a) Rubin ' s vase and (b) the duck – rabbit illusion

7.6 The principal pathways for saccade production

7.7 Platt and Glimcher ' s probabilistic cued-saccade task

7.8 Activity of an LIP neuron during the probability experiment

7.9 Platt and Glimcher ' s cued-saccade experiment, with stimulus and response held constant and the quantity of reward varied

7.10 Activity of an LIP neuron while a monkey makes its own choice compared to a behaviorally derived estimate of the value of the movement to the monkey

8.1 The architecture of a simple reflex agent

8.2 The architecture of a goal-based agent

8.3 The architecture of a learning agent

8.4 A version of the Wason selection task

8.5 A version of Griggs and Cox ' s deontic selection task

8.6 The evolutionary biologist W. D. Hamilton (1936 – 2000)

8.7 The ACT-R cognitive architecture

9.1 Luria ' s (1970) diagram of the functional organization of the brain

9.2 Map of the anatomy of the brain showing the four lobes and the Brodmann areas

9.3 A connectivity matrix for the visual system of the macaque monkey

9.4 An anatomical wiring diagram of the visual system of the macaque monkey

9.5 The results of single-neuron recordings of a mirror neuron in area F5 of the macaque inferior frontal cortex

9.6 Typical patterns of EEG waves, together with where/when they are typically found

9.7a Common experimental design for neurophysiological studies of attention

9.7b Example of the occipital ERPs recorded in a paradigm of this nature

9.7c Single-unit responses from area V4 in a similar paradigm

9.7d Single-unit responses from area V1 showing no effect of attention

9.8 Frontoparietal cortical network during peripheral visual attention

9.9 An illustration of a typical delayed saccade task

9.10 Peripheral attention versus spatial working memory versus saccadic eye movement across studies

10.1 The dual-route model of past-tense learning in English proposed by Steven Pinker and Alan Prince

10.2 Rumelhart and McClelland ' s model of past-tense acquisition

10.3 Performance data for Rumelhart and McClelland ' s model of past-tense learning

10.4 The network developed by Plunkett and Marchman to model children ' s learning of the past tense

10.5 A comparison of the errors made by Adam, a child studied by the psychologist Gary Marcus, and the Plunkett – Marchman neural network model of tense learning

10.6 A hierarchical cluster of similarity judgments, with nodes corresponding to clusters of stimuli more similar on average to each other than to objects in the nearest cluster

11.1 Schematic representation of the habituation and test

conditions in Baillargeon ′ s drawbridge experiments

11.2 Schematic representation of an experiment used to test infants ′ understanding of object boundaries and sensitivity to Spelke ′ s principle of cohesion (that surfaces lie on a single object if they are in contact)

11.3 Schematic representation of an experiment testing infants ′ understanding of the principle of contact (that only surfaces in contact can move together)

11.4 Schematic depiction of events that accord with, or violate, the continuity or solidity constraints

11.5 A series of inputs to the network as a barrier moves in front of a ball and then back to its original location

11.6 Recurrent network for learning to anticipate the future position of objects

11.7 A balance beam

11.8 The architecture of the McClelland and Jenkins network for the balance-beam problem

12.1 A decision tree illustrating a mortgage expert system

12.2 The first node on the decision tree for the tennis problem

12.3 The complete decision tree generated by the ID3 algorithm

12.4 A sample completed questionnaire used as input to an ID3-based expert system for diagnosing diseases in soybean crops

12.5 Different ways of distinguishing two groups in a database of examples

12.6 An illustration of hierarchical visual processing

12.7 Illustration of how an autoencoder compresses and then decompresses a signal

12.8 A move in the Google DeepMind challenge between AlphaGo and Lee Sedol in 2016

13.1 An example of metarepresentation

13.2 The general outlines of Leslie ' s model of pretend play

13.3 The task used by Baron-Cohen et al. to test for children ' s understanding of false belief

13.4 Illustration of the connection between pretend play and success on the false-belief task

13.5 Baron-Cohen ' s model of the mindreading system

13.6 A schematic version of standard simulationism

13.7 Schematic representation of brain regions associated with the attribution of mental states

13.8 Schematic overview of the frontoparietal mirror-neuron system (MNS) (*pink*) and its main visual input (*yellow*) in the human brain

14.1 A map of SHAKEY ' s physical environment

14.2 A photograph of SHAKEY the robot

14.3 The organizing principles of biorobotics — a highly interdisciplinary enterprise

14.4 The cricket ' s ears are on its front legs

14.5 A robot fish called WANDA

14.6 WANDA swimming upward

14.7 Another example of morphological computation: the robot hand designed by Hiroshi Yokoi

14.8 The Yokoi hand grasping two very different objects

14.9 Rodney Brooks ' s robot Allen, his first subsumption-architecture robot

14.10 The layers of Allen ' s subsumption architecture

14.11 The Nerd Herd, together with the pucks that they can pick up with their grippers

15.1 A typical priming experiment

15.2 Examples of deficits found in patients with left spatial neglect (damage to the right hemisphere of the brain)

15.3 D.B. ' s responses to pictures of animals presented in his blind field

15.4 An illustration of the two houses presented to P.S.

15.5 In this experiment, subjects were asked either to " post " a card into a slot or to rotate another handheld card to match the orientation of the slot

15.6 In the Ebbinghaus illusion, two circles are illusorily seen as differently sized, depending on what surrounds them

15.7 In the Norman and Shallice 1980 model (*top left*), conscious processing is involved in the supervisory attentional regulation, by the prefrontal cortices, of lower-level sensorimotor chains

15.8 The neural substrates of the global workspace

16.1 Facial expressions made by the subjects from New Guinea

16.2 The distinct physiological responses correlated with each basic emotion

16.3 The circumplex model of affect

16.4 Schematic representation of an emotional episode

16.5 Anatomical connections from the amygdala to brain areas directly involved in characteristic behavioral and physiological fear responses

16.6 The subjective experience of fear in S.M., as measured using standard self-report tools, in comparison with normal subjects

16.7 T2-weighted coronal images of the five subjects with UWD, with age in parentheses and cross-hairs indicating the location of the brain damage

16.8 Innate threat model and fear hypervigilance in UWD

16.9 The neural regions widely held to be implicated in the human fear network

16.10 Participants in the MRI scanner bed with the compartments

16.11 Two results from the Mobbs et al. (2010) neuroimaging study

16.12 Brain activity for the retreat-versus-approach contrast in Mobbs et al. (2010)

Tables

2.1 The three different levels that Marr identified for explaining information-processing systems

4.1 Syntax and semantics in the predicate calculus

8.1 Comparing the symbolic and subsymbolic dimensions of knowledge representation in the hybrid ACT-R architecture

9.1 Comparing techniques for studying connectivity in the brain

10.1 The stages of past-tense learning according to verb type

14.1 SHAKEY ' s five levels

14.2 SHAKEY ' s intermediate-level routines

14.3 The five basis behaviors programmed into Matari ć' s Nerd Herd robots

16.1 Results of facial expression recognition in New Guinea adult subjects

16.2 Classifying different affective phenomena by typical duration and role

16.3 The dimensionality of emotion according to an influential version of cognitive-appraisal theory

16.4 Tools for studying different components of an emotional episode

16.5 Measures of fear in animal studies compared to the *DSM*

criteria for generalized anxiety

Boxes

2.1 A Conversation with ELIZA

3.1 What Does Each Lobe Do?

3.2 Brain Vocabulary

4.1 Defining Sentences in Propositional Logic

6.1 Basins of Attraction in State Space

7.1 Basics of the Probability Calculus

7.2 Deriving Bayes ′ s Rule

8.1 The Prisoner ′ s Dilemma

Preface

About This Book

Few things are more fascinating than the human mind — and few are more difficult to understand. Cognitive science is the enterprise of trying to make sense of this most complex and baffling natural phenomenon.

The very things that make cognitive science so fascinating make it very difficult to study and to teach. Many different disciplines study the mind. Neuroscientists study the mind ' s biological machinery. Psychologists directly study mental processes, such as perception and decision-making. Computer scientists explore how those processes can be simulated and modeled in computers. Evolutionary biologists and anthropologists speculate about how the mind evolved. In fact, very few academic areas are not relevant to the study of the mind in some way. The job of cognitive science is to provide a framework for bringing all these different perspectives together.

The enormous range of information out there about the mind can be overwhelming, both for students and for instructors. Different textbooks have approached this challenge in different ways.

Some textbooks have concentrated on being as comprehensive as possible, with a chapter covering key ideas in each of the relevant

disciplines — a chapter on psychology, a chapter on neuroscience, and so on. These books are often written by committees — with each chapter written by an expert in the relevant field. These books can be very valuable, but they really give an introduction to the cognitive sciences (in the plural) rather than to cognitive science as an interdisciplinary enterprise.

Other textbook writers take a much more selective approach, introducing cognitive science from the perspective of the disciplines that they know best — from the perspective of philosophy, for example, or of computer science. Again, I have learned much from these books, and they can be very helpful. But I am convinced that students and instructors need something more general.

This book aims for a balance between these two extremes. Cognitive science has its own problems and its own theories. The book is organized around these. They are all ways of working out the fundamental idea at the heart of cognitive science — which is that the mind is an information processor. What makes cognitive science so rich is that this single basic idea can be (and has been) worked out in many different ways. In presenting these different models of the mind as an information processor, I have tried to select as wide a range of examples as possible to give students a sense of cognitive science ′s breadth and range.

About the Fourth Edition

This fourth edition of *Cognitive Science: An Introduction to the Science of the Mind* has been significantly revised in response to suggestions from

users of the three previous editions and other experts commissioned by Cambridge University Press.

The major addition is Chapter 16 , coauthored with Dong An, formerly a PhD student at Texas A&M and now on the faculty at Zhejiang University in China. This new chapter introduces the emerging area of affective science. Affective science studies the emotions and other affective states. It uses some of the techniques and tools discussed in Part II, but it has also provided fertile ground for applying new techniques, particularly in the realm of genetics. Chapter 16 on affective science complements Chapter 15 on consciousness: both chapters explore aspects of the mind that traditional cognitive science has neglected. Chapter 15 has itself been reorganized for this edition. I have, at the instigation of David Rosenthal, introduced discussion of the higher-order theory of consciousness.

I have consolidated what previously, in the third edition, was two separate chapters on mindreading (" Exploring Mindreading " and " Mindreading: Advanced Topics ") into what is now Chapter 13 . Most of the main points from the early chapters have been preserved, but the presentation is now smoother, and some of the intricacies have gone.

Every chapter has been reviewed for concision and readability, and all the annotated bibliographies have been updated to incorporate significant new publications through 2021. The end result is a book that is slightly shorter than its predecessor, which I anticipate will allow Cambridge University Press to continue to offer competitive pricing.

How the Book Is Organized

This book is organized into three parts.

Part I: Historical Landmarks

Cognitive science has evolved considerably in its short life. Priorities have changed as new methods have emerged — and some fundamental theoretical assumptions have changed with them. The three chapters in Part I introduce students to some of the highlights in the history of cognitive science. Each chapter is organized around key discoveries and/or theoretical advances.

Part II: Models and Tools

Part II sets out the main models and tools that cognitive scientists can bring to bear to understand cognition and the mind.

The first model, discussed in Chapter 4 , is associated with the physical symbol system hypothesis originally developed by computer scientists Allen Newell and Herbert Simon. According to the physical symbol system hypothesis, all information processing involves the manipulation of physical structures that function as symbols. For the first decades of cognitive science, the physical symbol systems hypothesis was, as the late Jerry Fodor famously put it, the " only game in town. " In the 1980s and 1990s, connectionist and neural network modelers developed an alternative, derived from models of artificial neurons in computational neuroscience and connectionist artificial intelligence. Chapter 5 explores the motivation for this approach and introduces some of the key concepts.

Another set of models and tools that can be used to study the mind derives from dynamical systems theory and is introduced and discussed in Chapter 6 . Bayesian approaches to modeling the mind have also gained currency. As explained in Chapter 7 , these approaches treat the mind as a predictive, hypothesis-testing machine and have been used both to study the mind as a whole and to model the activity of individual brain areas and populations of neurons.

One of the key ideas of cognitive science is that the mind is modular (that some, or all, information processing is carried out by specialized modules). Chapter 8 explores different ways of developing this basic idea, including the radical claim, proposed by evolutionary psychologists, that

the mind is simply a collection of specialized modules, with no nonspecialized processing at all. Theoretical discussions of modularity are complemented by experimental techniques for studying the organization of the mind. Chapter 9 surveys the cognitive scientist ′ s toolkit in this regard, focusing in particular on different types of brain mapping.

Part III: Applications

The eight chapters in this part are more applied than those in Part II. They explore different ways in which the models and tools introduced in Part II can be used to give accounts of particular cognitive phenomena.

Chapter 10 considers language learning. Many cognitive scientists have thought that language learning is a process of learning explicit rules with a significant innate component. But we explore both neural network and Bayesian models that illustrate alternative ways of thinking about how children can learn languages. In Chapter 11 , we turn to models of how children learn about the basic structure of the physical world (how they acquire what is often called a *folk physics*). Here, too, we see the power of neural network models.

One of the most significant recent developments of neural network models has been the explosive growth of deep-learning algorithms, which have made possible impressive advances in areas long thought to be major challenges for artificial intelligence, such as image recognition, machine translation, and games of strategy, such as Go. These are covered in Chapter 12 .

Chapters 13 and 14 illustrate how theoretical, methodological, and experimental issues can come together. Chapter 13 works through an issue that has received much attention in contemporary cognitive science — the issue of whether there is a dedicated cognitive-system response for our understanding of other people (the so-called *mindreading system*). Chapter 14 explores the situated cognition movement and related developments in

robotics, particularly behavior-based robotics and biologically inspired robotics.

In Chapter 15 , we look at recent developments in the cognitive science of consciousness — a fast-moving and exciting area that raises fundamental questions about possible limits to what can be understood through the tools and techniques of cognitive science. And then finally, in Chapter 16 , we examine the emerging area of affective science, which uses multiple techniques to study emotions and other affective states in human and nonhuman animals. The book ends with Chapter 17 , which briefly sketches out a selection of exciting new directions in cognitive science.

Using This Book in Courses

This book has been designed to serve as a self-contained text for a single-semester (twelve to fifteen weeks) introductory course on cognitive science. Students taking this course may have taken introductory courses in psychology and/or philosophy, but no particular prerequisites are assumed. All the necessary background is provided for a course at the freshman or sophomore level (first or second year). The book could also be used for a more advanced introductory course at the junior or senior level (third or fourth year). In this case, the instructor would most likely want to supplement the book with additional readings. There are suggestions on the instructor website (see " Course Website " section).

Text Features

I have tried to make this book as user-friendly as possible. Key text features include the following:

- **Chapter overviews.** Each chapter begins with an overview to orient the reader.
- **Exercises.** These have been inserted at various points within each chapter. They are placed in the flow of the text to encourage the reader to take a break from reading and engage with the material. They are typically straightforward, but for a few, I have placed

suggested solutions on the instructor website (see " Course Website " section).

- **Boxes.** Boxes have been included to provide further information about the theories and research discussed in the text. Readers are encouraged to work through these, but the material is not essential to the flow of the text.
- **Summaries, checklists, and further reading.** These can be found at the end of each chapter. The summary provides a short overview of the chapter. The checklists allow students to review the key points of the chapter and also serve as a reference point for instructors. Suggestions of additional books and articles are provided to guide students ' further reading on the topics covered in the chapter.

Course Website

A course website accompanies the book. It can be found at www.cambridge.org/bermudez4e . The **publicly accessible section of the website** contains the following:

- Discussion questions for each chapter that students can use to check their understanding
- Sample syllabi for courses of different lengths and different levels
- Links to useful learning resources, videos, and experimental demonstrations

- Links to online versions of relevant papers and online discussions for each chapter

Instructors and other authorized users can access a **password-protected section of the website** . This contains the following:

- PowerPoint slides for each chapter
- Electronic versions of figures from the text
- Suggested solutions for the more challenging exercises and problems
- Student review questions
- A bank of questions that instructors can use for assessment (the questions are formatted to be uploaded into Respondus 4.0 for compatibility with many learning management systems or imported directly into Blackboard)

The website is a work in progress. Students and instructors are welcome to contact me with suggestions, revisions, and comments. Contact details are on the website.

Acknowledgments

Acknowledgments for the First Edition

Many friends and colleagues associated with the Philosophy – Neuroscience – Psychology program at Washington University in St. Louis have commented on sections of this book. I would particularly like to thank Maurizio Corbetta, Frederick Eberhardt, David Kaplan, Clare Palmer, Gualtiero Piccinnini, Marc Raichle, Philip Robbins, David Van Essen, and Jeff Zacks. Josef Perner kindly read a draft of Chapter 12 .

I have benefited from the comments of many referees while working on this project. Most remain anonymous, but some have revealed their identity. My thanks to Kirsten Andrews, Gary Bradshaw, Rob Goldstone, Paul Humphreys, and Michael Spivey.

Drafts of this textbook have been used four times to teach PNP 200 Introduction to Cognitive Science here at Washington University in St. Louis — twice by me and once each by David Kaplan and Jake Beck. Feedback from students both inside and outside the classroom was extremely useful. I hope that other instructors who use this text have equally motivated and enthusiastic classes. I would like to record my thanks to the teaching assistants who have worked with me on this course: Juan Montaña, Tim Oakberg, Adam Shriver, and Isaac Wiegman. And also to

Kimberly Mount, the PNP administrative assistant, whose help with the figures and preparing the manuscript is greatly appreciated.

A number of students from my spring 2009 PNP 200 class contributed to the Glossary. It was a pleasure to work with Olivia Frosch, Katie Lewis, Juan Manfredi, Eric Potter, and Katie Sadow.

Work on this book has been made much easier by the efforts of the psychology textbook team at Cambridge University Press — Raihanah Begum, Catherine Flack, Hetty Reid, Sarah Wightman, and Rachel Willsher (as well as to Andy Peart, who signed this book up but has since moved on). They have been very patient and very helpful. My thanks also to Anna Oxbury for her editing and to Liz Davey for coordinating the production process.

Acknowledgments for the Second Edition

I am very grateful to my colleagues in the Office of the Dean of Liberal Arts at Texas A&M University, particularly my administrative assistant Connie Davenport, for helping me to carve out time to work on the second edition of the textbook. T. J. Kasperbauer has been an excellent research assistant, providing numerous improvements to the text and supporting resources and helping me greatly with his deep knowledge of cognitive science. It has been a pleasure to work once again with Hetty Marx and Carrie Parkinson at Cambridge University Press. I particularly appreciate their work in gathering feedback on the first edition. Thanks again to Anna Oxbury for her copyediting skills.

Acknowledgments for the Third Edition

Part of the work on the third edition was carried out during a period of leave in the 2018 – 2019 academic year. I am grateful to the Department of Philosophy at Texas A&M University for allowing me to take a year off from teaching and administrative duties. At Cambridge University Press, I worked initially with Janka Romero and then subsequently with Stephen Acerra and Lisa Pinto, all of whom were very supportive and worked hard to get me helpful feedback on earlier editions. I am particularly grateful to the many reviewers who gave detailed comments and suggestions for improvement. Finally, I ′ d like to record my thanks to Dong An for her help with the online resources.

Acknowledgments for the Fourth Edition

It has been a pleasure to work with Stephen Acerra, Stefanie Seaton, and the team at Cambridge University Press, who obtained much valuable feedback on the third edition and helped steer the new edition into production. I am particularly grateful to the nearly thirty reviewers who sent in comments on the third edition. The book has been enormously improved as a result.

As I was completing work on the new edition, I found myself unexpectedly taking on heavy administrative duties at Texas A&M. I am very grateful to provost Tim Scott for providing funding for a graduate research assistant to keep the project on track. Dong Yu (David) Yang

greatly assisted me in updating the annotated bibliographies and online resources.

Finally, I am delighted that Dong An, who assisted me with the third edition while still a PhD student at A&M, signed up as a coauthor for the new Chapter 16 on emotions. Dong is now on the faculty at Zhejiang University, one of the leading universities in China, and I am grateful for her expert guidance on a topic that she knows much better than I do.

Image credits: Part I: © SciePro / iStock / Getty Images Plus; Part II: © Andrzej Wojcicki / Science Photo Library / Getty Images; Part III © Paper Boat Creative / DigitalVision / Getty Images.

Introduction:

The Challenge of Cognitive Science



Overview

Cognitive Science: An Interdisciplinary Endeavor

Levels of Explanation: The Contrast between Psychology and Neuroscience

- How Psychology Is Organized
- How Neuroscience Is Organized

The Challenge of Cognitive Science

- Three Dimensions of Variation
- The Space of Cognitive Science

Overview

Cognitive science draws upon the tools and techniques of many different disciplines, including psychology, philosophy, linguistics, computer science, neuroscience, and mathematical logic. It is a fundamentally *interdisciplinary activity* . This basic fact raises important and fundamental questions. What do all these disciplines have in common? How can they all come together to form a distinctive area of inquiry?

The aim of this Introduction is to give you a sense of the scope and range of cognitive science, setting the framework for more detailed study in subsequent chapters. We will explore the idea that the different disciplines in cognitive science each study different levels of organization in the mind and the nervous system. In particular, we will see how the brain can be studied at many different levels, from the level of the molecule upward. The Introduction ends with a description (and illustration) of what I call the *space of cognitive science* .

Cognitive Science: An Interdisciplinary Endeavor

The hexagonal diagram in Figure 0.1 is one of the most famous images in cognitive science. It comes from the 1978 report on the state of the art in cognitive science commissioned by the Sloan Foundation and written by a group of leading scholars. The diagram is intended to illustrate the interdisciplinary nature of cognitive science. The lines on the diagram indicate the academic disciplines that the authors saw as integral parts of cognitive science, together with the connections between disciplines particularly relevant to the study of mind and cognition.

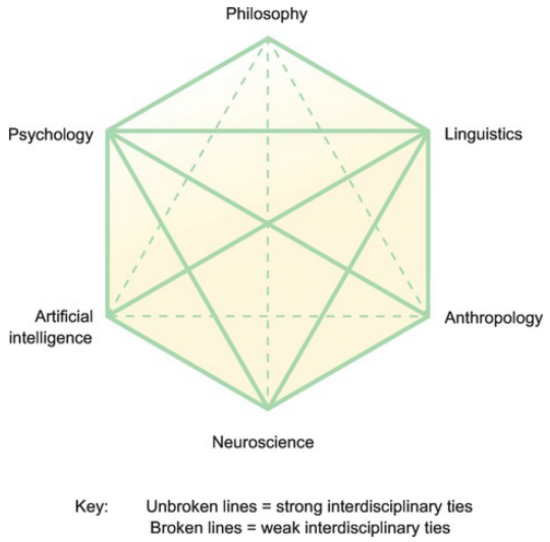


Figure 0.1 Connections among the cognitive sciences, as depicted in the Sloan Foundation ′ s 1978 report. Unbroken lines indicate strong interdisciplinary links, whereas broken lines indicate weaker links.

(Adapted from Gardner 1985.)

For the authors of the Sloan report, cognitive science is an amalgamation of philosophy, psychology, linguistics, anthropology, neuroscience, and artificial intelligence. Each of the six disciplines brings with it different techniques, tools, and frameworks for thinking about the mind. Each of them studies the mind from different perspectives and at different levels. Whereas linguists, for example, develop abstract models of linguistic *competence* (the abstract structure of language), psychologists of language are interested in the mechanisms that make possible the *performance* of language users. Whereas neuroscientists study the details of how the brain works, computer scientists abstract away from those details to explore computer models and simulations of human cognitive abilities. Anthropologists are interested in the social dimensions of cognition, as well as how cognition varies across cultures. Philosophers, in contrast, are

typically interested in very abstract models of how the mind is realized by the brain.

Some of the connections identified in the diagram were judged stronger than others. These are marked with a solid line. The weaker connections are marked with a broken line. At least one of the connections that were judged weak in 1978 has now become a thriving subdiscipline in its own right. Some philosophers impressed by the potential for fruitful dialog between philosophy and neuroscience have taken to calling themselves *neurophilosophers* , after the title of a very influential book by Patricia Churchland published in 1986.

Miller ′ s (2003) account of how the Sloan report was written is both disarming and telling. ‘‘ The committee met once, in Kansas City. It quickly became apparent that everyone knew his own field and had heard of two or three interesting findings in other fields. After hours of discussion, experts in discipline X grew unwilling to make any judgments about discipline Y, and so forth. In the end, they did what they were competent to do: each summarized his or her own field and the editors — Samuel Jay Keyser, Edward Walker and myself — patched together a report ″ (Miller 2003 : 143). This may be how reports get written, but it is not a very good model for an interdisciplinary enterprise such as cognitive science.

In fact, the hexagon as a whole is not a very good model for cognitive science. Even if we take seriously the lines that mark connections between the disciplines of cognitive science, the hexagon gives no sense of a unified intellectual enterprise. It gives no sense, that is, of something that is more than a composite of ‘‘ traditional ″ disciplines such as philosophy and psychology. There are many different schools of philosophy and many different specializations within psychology, but there are certain things that

bind together philosophers as a group and psychologists as a group, irrespective of their school and specialization. For philosophers (particularly in the so-called *analytic* tradition, the tradition most relevant to cognitive science), the unity of their discipline comes from certain problems that are standardly accepted as philosophical, together with a commitment to rigorous argument and analysis. The unity of psychology comes, in contrast, from a shared set of experimental techniques and paradigms. Is there anything that can provide a similar unity for cognitive science?

One of the main claims of this textbook is that cognitive science is indeed a unified enterprise. It has its own distinctive problems, its own distinctive techniques, and its own distinctive explanatory frameworks. We will be studying all of these in this book. First, though, we need to get a better picture of the range and scope of the enterprise. In the rest of this Introduction, I ′ ll use psychology and neuroscience as examples to give you a sense of the overall space of cognitive science.

Levels of Explanation: The Contrast between Psychology and Neuroscience

Neuroscience occupies one pole of the Sloan report ′ s hexagonal figure, and it was not viewed as very central to cognitive science by the authors of the report. The report was written before the " turn to the brain " that we will look at in Chapter 3 , and its focus reflected the contemporary focus on computer science, psychology, and linguistics as the core disciplines of cognitive science. Moreover, the authors of the report treated neuroscience as a unitary discipline, on a par with anthropology, psychology, and other

more traditional academic disciplines. The explosion of research into what became known as *cognitive neuroscience* has since corrected both of these assumptions. Most cognitive scientists place the study of the brain firmly at the heart of cognitive science. And it is becoming very clear that neuroscience is itself a massively interdisciplinary field.

How Psychology Is Organized

One way of thinking about what distinguishes neuroscience from, say, psychology is through the idea of levels . I am talking here about what is sometimes called *scientific psychology* (psychology as it is taught and studied in university departments), as opposed, for example, to humanistic psychology, self-help psychology, and much of what is routinely classified as psychology in bookstores. But even narrowing it down like this, there are many different subfields of psychology.

A quick look at the courses on offer in any reputable psychology department will find courses in cognitive psychology, social psychology, abnormal psychology, personality psychology, psychology of language, and so on. It is normal for research psychologists to specialize in, at most, one or two of these fields. Nonetheless, most psychologists think that psychology is a single academic discipline. This is partly because there is a continuity of methodology across the different specializations and subfields. Students in psychology are typically required to take a course in research methods. Such courses cover basic principles of experimental design, hypothesis formation and testing, and data analysis that are common to all branches of psychology.

Equally important, however, is the fact that many of these branches of psychology operate at the same level. The data from which they begin are data about cognitive performance and behavior at the level of the whole organism (I am talking about the whole organism to make clear that these ideas extend to nonhuman organisms, as studied in comparative psychology).

The basic *explananda* (the things that are to be explained) in psychology are people ′ s psychological capacities, which include both cognitive and emotional capacities. The organization of psychology into different subfields is a function of the fact that there are many different types of cognitive and emotional capacities.

Within cognitive psychology, for example, what psychologists are trying to explain is the organism ′ s capacities for perception, memory, attention, and so on. Controlled experiments and correlational studies are used to delimit and describe those capacities so that psychologists know exactly what it is that needs to be explained.

Likewise, social psychologists study the capacities involved in social understanding and social interactions. They are interested, for example, in social influences on behavior, on how we respond to social cues, and on how our thoughts and feelings are influenced by the presence of others. Personality psychologists study the traits and patterns of behavior that make up what we think of as a person ′ s character. And so on.

If we were to map out some of the principal subfields in scientific psychology, it would look something like Figure 0.2 . The diagram is intended to show that the different sub-branches all study different aspects of mind and behavior at the level of the organism.

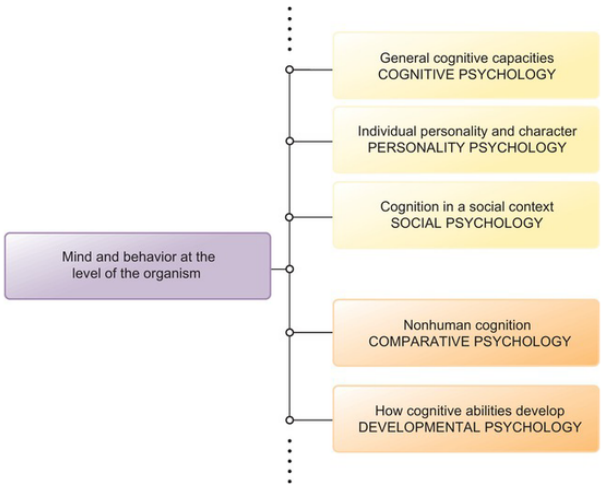


Figure 0.2 Some of the principal branches of scientific psychology.

How Neuroscience Is Organized

Things are very different in neuroscience. There are many branches of neuroscience, but they are not related in the same way. The organization of neuroscience into branches closely follows the different levels of organization in the brain and the central nervous system. These levels of organization are illustrated in Figure 0.3 , drawn from Gordon Shepherd ′ s 1994 textbook *Neurobiology* .

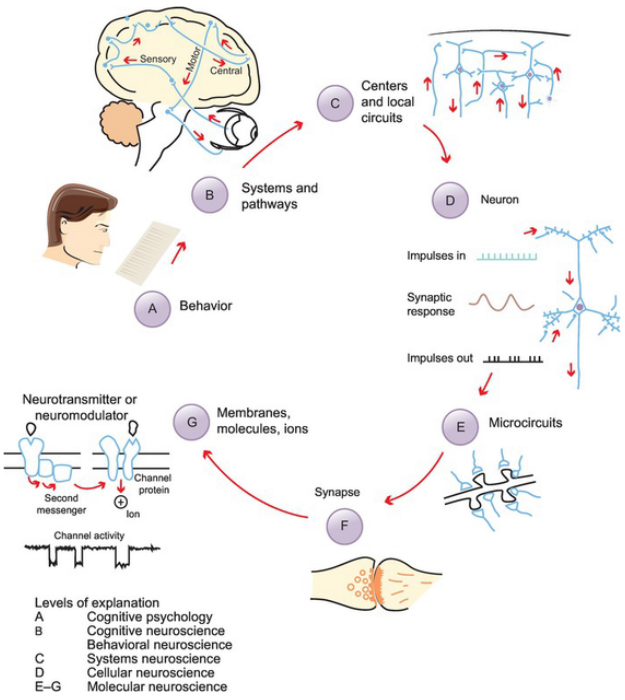


Figure 0.3 Levels of organization and levels of explanation in the nervous system.

(Adapted from Shepherd 1994 .)

You may have come across references to areas in the brain such as the primary visual cortex or the hippocampus, for example. And you may have encountered talk of neural pathways connecting different areas in the brain. Located at levels A and B in Shepherd ′ s diagram, these are the highest levels of neural organization, corresponding most closely to cognitive activities that we all perform. The primary visual cortex, for example, is responsible for coding the basic features of visual information coming from the retina. It is sensitive to orientation, motion, speed, direction, and so on. The hippocampus, in contrast, is thought to be responsible for key aspects of memory.

Activity at this top level of organization is the result of activity at lower levels of organization. In Shepherd ′ s diagram, this takes us to levels

C and E — the level of centers, local circuits, and microcircuits. Somehow, the collective activity of populations of neurons codes certain types of information about objects in a way that organizes and coordinates the information carried by individual neurons. These populations of neurons are the local circuits in Shepherd ′ s diagram.

What happens in populations of neurons is ultimately determined by the behavior of individual neurons. But neurons are not the most basic level of organization in the nervous system. In order to understand how neurons work, we need to understand how they communicate. This brings us to Shepherd ′ s level F because neurons communicate across synapses. Most synapses are chemical, but some are electrical. The chemical synapses work through the transmission of neurochemicals (*neurotransmitters*). These neurotransmitters are activated by the arrival of an electrical signal (the *action potential*). The propagation of neurotransmitters works the way it does because of the molecular properties of the synaptic membrane — properties that are ultimately genetically determined. With this, we arrive at level G in Shepherd ′ s diagram .

The point of this whistle-stop tour through the levels of organization in the brain is that the subfields of neuroscience map very closely onto the different levels of organization in the brain. At the top level, we have cognitive neuroscience and behavioral neuroscience, which study the large-scale organization of the brain circuits deployed in high-level cognitive activities. These operate at what I termed, in discussing the subfields of psychology, the level of the whole organism. Systems neuroscience, in contrast, investigates the functioning of neural systems, such as the visual system. The bridge between the activity of neural systems and the activity of individual neurons is one of the central topics in computational

neuroscience, whereas cellular and molecular neuroscience deal with the fundamental biological properties of neurons.

Different branches of neuroscience (and cognitive science in general) employ tools appropriate to the level of organization at which they are studying the brain. These tools and techniques vary in what neuroscientists call their *temporal* and *spatial resolution* . That is, they vary in the scale on which they give precise measurements (spatial resolution) and the time intervals to which they are sensitive (temporal resolution).

Some of the important variations are depicted in Figure 0.4 . We will explore the differences between these different tools and technologies in much more detail in later chapters (particularly Chapter 9).

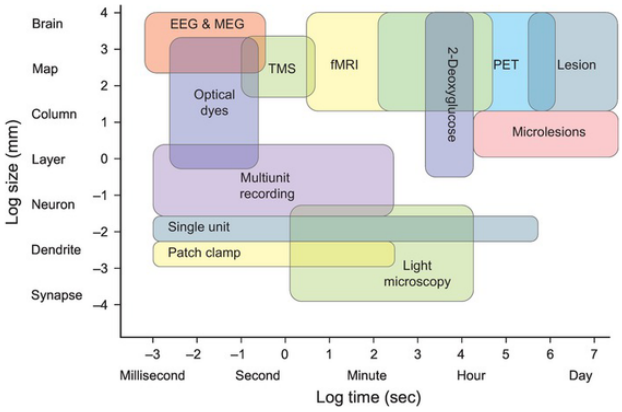


Figure 0.4 The spatial and temporal resolution of different tools and techniques in neuroscience. Time is on the *x* -axis, and size is on the *y* -axis.

(Adapted from Baars and Gage 2010 .)

The Challenge of Cognitive Science

This section explores these basic ideas of levels of organization, levels of resolution, and levels of explanation further to give a picture of what I call the *space of cognitive science* .

Three Dimensions of Variation

Cognitive science draws upon a large number of potentially relevant fields and subfields. Those fields and subfields differ from each other along three dimensions.

One dimension of variation is illustrated by the subfields of neuroscience. Neuroscience studies the brain at many different levels. These levels are organized into a vertical hierarchy that corresponds to the different levels of organization in the nervous system.

A second dimension of variation comes with the different techniques and tools that cognitive scientists can employ. As illustrated in Figure 0.4 , these tools vary both in spatial and in temporal resolution. Some tools, such as positron emission tomography (PET) and functional magnetic resonance imaging (fMRI), give accurate measurements at the level of individual brain areas. Others, such as microelectrode recording, give accurate measurements at the level of individual neurons (or small populations of neurons).

The third dimension of variation is exemplified by the different subfields of psychology. Most of psychology operates at Shepherd ′ s level A. The different areas of psychology set out to explore, map, describe, and explain the cognitive abilities making possible the myriad things that human beings do and say.

The Space of Cognitive Science

The different parts of cognitive science are distributed, therefore, across a three-dimensional space, as illustrated in Figure 0.5 .

- The *x* -axis marks the different cognitive domains that are being studied.
- The *y* -axis marks the different tools that might be employed (ordered roughly in terms of their degree of spatial resolution).
- The *z* -axis marks the different levels of organization at which cognition is studied.

This three-dimensional diagram is a more accurate representation of where cognitive science stands in the early years of the twenty-first century than the two-dimensional hexagon proposed by the authors of the Sloan report (although the hexagon may well have been an adequate picture of how things stood at the end of the 1970s).

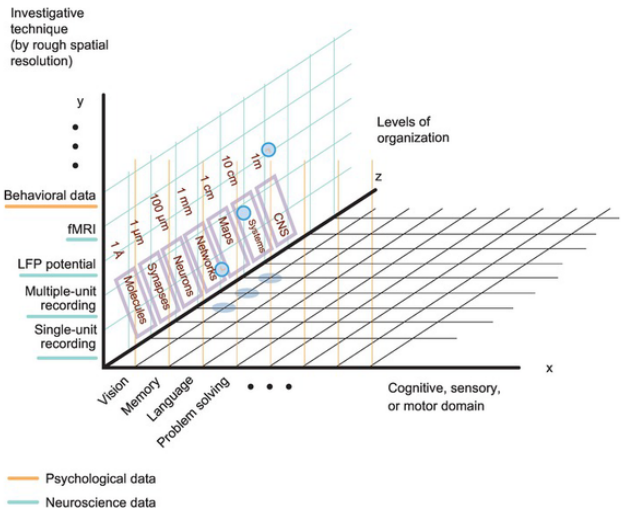


Figure 0.5 The " space " of contemporary cognitive science.

A good way of thinking about cognitive science is as setting out to provide a unified account of cognition that draws upon and integrates the whole space. Cognitive science is more than just the sum of its parts. The aim of cognitive science as an intellectual enterprise is to provide a framework that makes explicit the common ground between all the different academic disciplines that study the mind and that shows how they are related to each other.

You can think of the analogy with physics. Many theoretical physicists think that the ultimate goal of physics is to provide a unified " theory of everything. " So, too (in this way of thinking about cognitive science), is it the mission of cognitive science to provide a unified " theory of cognition. "

Parts II and III explore the principal theories of cognition in cognitive science and show how they can be applied to explain different aspects of cognition. First, though, we turn to an overview of some of the key historical landmarks in the emergence and subsequent development of cognitive science. That will occupy the three chapters of Part I. These chapters should put flesh on the bones of the general picture sketched out in this introduction.

Further Reading

Historical background on the Sloan report can be found in Gardner 1985 and Miller 2003 (available in the online resources). The report itself was never published. A very useful basic introduction to levels of organization and structure in the nervous system is Chapter 2 of Churchland and

Sejnowski 1992 . For more detail, a classic neuroscience textbook is Kandel et al. 2012 . Stein and Stoodley 2006 and Purves et al. 2011 are alternatives. Craver 2007 discusses the interplay between different levels of explanation in the neuroscience of memory. Piccinini and Craver 2011 is a more general discussion; also see Bickle 2006 and Sullivan 2009 .

Part I



Historical Landmarks

Chapter 1

The Prehistory of Cognitive Science

◈

Overview

1.1 The Reaction against Behaviorism in Psychology

Learning without Reinforcement: Tolman and Honzik, "' Insight ' in Rats " (1930)

Cognitive Maps in Rats? Tolman et al., " Studies in Spatial Learning " (1946)

Plans and Complex Behaviors: Lashley, " The Problem of Serial Order in Behavior " (1951)

1.2 The Theory of Computation and the Idea of an Algorithm

Algorithms and Turing Machines: Turing, " On Computable Numbers, with an Application to the Decision Problem " (1936 – 1937)

1.3 Linguistics and the Formal Analysis of Language

The Structure of Language: Chomsky ' s *Syntactic Structures* (1957)

1.4 Information-Processing Models in Psychology

How Much Information Can We Handle? George Miller ′ s “ The Magical Number Seven, Plus or Minus Two ” (1956)
The Flow of Information: Donald Broadbent ′ s “ The Role of Auditory Localization in Attention and Memory Span ” (1954) and *Perception and Communication* (1958)
1.5 Connections and Points of Contact

Overview

By the late 1970s, cognitive science was an established part of the intellectual landscape, crystallized around a basic set of problems, techniques, and theoretical assumptions. These came from many different disciplines and areas, and many had been around for a fairly long time. What was new was the idea of putting them together as a way of studying the mind.

A good way to understand cognitive science is to try to think your way back to how things might have looked to its early pioneers. They were exploring a landscape in which certain regions were well mapped and well understood but where there were no standard ways of getting from one region to another. An important part of what they did was to show how these different regions could be connected in order to create an interdisciplinary science of the mind.

This chapter explores the *prehistory* of cognitive science in the 1930s, 1940s, and 1950s. It looks at some of the basic ideas and currents of thought that, in retrospect, we can see as feeding into what came to be known as *cognitive science* . As we shall see in more detail later on, *the guiding idea of cognitive science is that mental operations involve*

processing information , and hence, we can study how the mind works by studying how information is processed. This basic idea of the mind as an information processor has a number of very specific roots in areas that seem initially to have little in common. The prehistory of cognitive science involves parallel, and largely independent, developments in psychology, linguistics, and mathematical logic. We will be looking at four of these developments:

- The reaction against behaviorism in psychology (Section 1.1)
- The idea of algorithmic computation in mathematical logic (Section 1.2)
- The emergence of linguistics as the formal analysis of language (Section 1.3)
- The emergence of information-processing models in psychology (Section 1.4)

In concentrating on these four developments, we will be passing over other important influences, such as neuroscience and neuropsychology. This is because until quite recently, the direct study of the brain had a relatively minor role to play in cognitive science.

Almost all cognitive scientists are convinced that in some fundamental sense, the mind just is the brain, so everything that happens in the mind is happening in the brain. Few, if any, cognitive scientists are *dualists* who think that the mind and the brain are two separate and distinct things. But for a long time in the history of cognitive science, it was widely held that we are better off studying the mind by abstracting away from the details of what is going on in the brain. This changed only with the emergence in the

1970s and 1980s of new technologies for studying neural activity and of new ways of modeling cognitive abilities — as we will see in Chapter 3 .

1.1 The Reaction against Behaviorism in Psychology

Behaviorism was (and in some quarters remains) influential in psychology. It takes many different forms, but all share the basic assumption that psychologists should confine themselves to studying observable phenomena and measurable behavior. Instead of speculating about unobservable mental states, psychologists should instead focus on nonpsychological mechanisms linking particular stimuli with particular responses. These mechanisms are the product of conditioning. For examples of conditioning, think of Pavlov ' s dogs being conditioned to salivate at the sound of the bell or the rewards/punishments that animal trainers use to encourage/discourage certain types of behavior.

For behaviorists, psychology is really the science of behavior, and so cognitive science could not even get started until behaviorism ceased to be dominant. Psychology ' s move from behaviorism was a lengthy and drawn-out process (and some would say that it has not yet been completed). We can appreciate some of the ideas that proved important for the later development of cognitive science by looking at three landmark papers. Each was an important statement of the idea that various types of behavior could not be explained in terms of stimulus – response mechanisms. Instead, psychologists need to think about organisms as storing and

processing information about their environment, rather than as responding mechanically to reinforcers and stimuli.

Learning without Reinforcement: Tolman and Honzik, "' Insight ' in Rats " (1930)

Edward Tolman (1886 – 1959) was a behavioral psychologist studying problem solving and learning in rats (among other things). As with most psychologists of the time, he started off with two standard behaviorist assumptions about learning. The first assumption is that all learning is the result of *conditioning.* The second assumption is that conditioning depends on the processes of *association* and *reinforcement.*

To get a fix on these two assumptions, consider a rat in what is known as a *Skinner box* , after the celebrated behaviorist B. F. Skinner. A typical Skinner box is illustrated in Figure 1.1 . The rat receives a reward each time it behaves in a particular way (pressing a lever, for example, or pushing a button). The reward *reinforces* the behavior. This means that the association between the behavior and the reward is strengthened, and it becomes more likely that the rat will perform the behavior again. The rat becomes *conditioned* to perform the behavior.

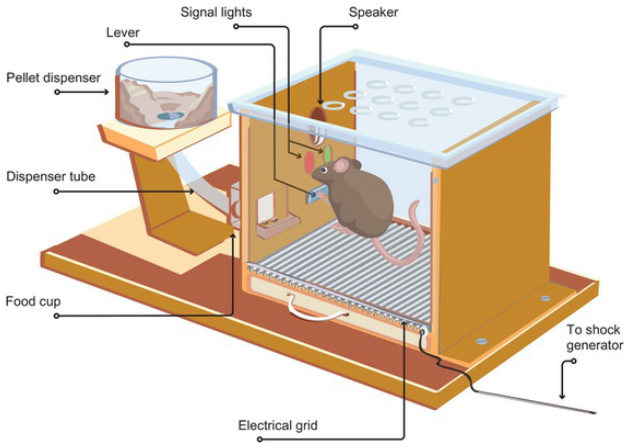


Figure 1.1 A rat in a Skinner box. The rat has a response lever controlling the delivery of food, as well as devices allowing different types of stimuli to be produced.

The basic idea of behaviorism is that all learning is either reinforcement learning of this general type or the even simpler form of associative learning often called *classical conditioning* .

In classical conditioning, what is strengthened is the association between a *conditioned stimulus* (such as the typically neutral sound of a bell ringing) and an *unconditioned stimulus* (such as the presentation of food). The unconditioned stimulus is *not* neutral for the organism and typically provokes a behavioral response, such as salivation. During classical conditioning, the association between the conditioned stimulus and the unconditioned stimulus is strengthened, which eventually leads the organism to produce the unconditioned response to the conditioned stimulus alone, without the presence of the unconditioned stimulus. The most famous example is Pavlov ' s dogs, which were conditioned to salivate to the sound of a bell by the simple technique of using the bell to signal the arrival of food.

So, it is a basic principle of behaviorism that all learning, whether by rats or by human beings, takes place through processes of reinforcement and conditioning. What the studies reported by Tolman and Honzik in 1930 seemed to show, however, is that this is not true even for rats.

Tolman and Honzik were interested in how rats learned to navigate mazes. They ran three groups of rats through a maze of the type illustrated in Figure 1.2 . The first group received a reward each time they successfully ran the maze. The second group never received a reward. The third group was unrewarded for the first ten days and then began to be rewarded.

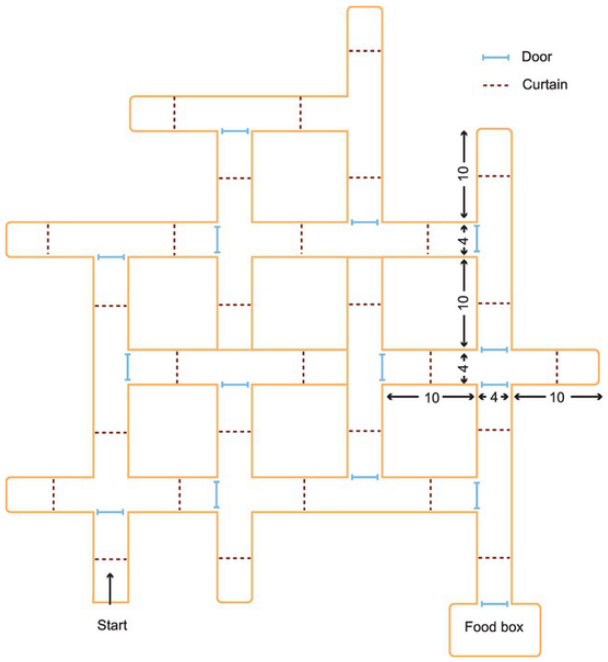


Figure 1.2 A fourteen-unit T-Alley maze (measurements in inches). Note the blocked passages and dead ends.

(Adapted from Elliott 1928 .)

As behaviorism predicted, the rewarded rats quickly learned to run the maze, whereas both groups of unrewarded rats simply wandered around aimlessly. The striking fact, however, was that when the third group of rats

started to receive rewards, they learned to run the maze far more quickly than the first group had.

Tolman and Honzik argued that the rats must have been learning about the layout of the maze during the period when they were not being rewarded. This type of *latent learning* seemed to show that reinforcement was not necessary for learning and that the rats must have been picking up and storing information about the layout of the maze when they were wandering around it, even though there was no reward and hence no reinforcement. They were later able to use this information to navigate the maze.

Exercise 1.1 Explain in your own words why latent learning seems to be incompatible with the two basic assumptions of behaviorism.

Suppose, then, that organisms are capable of latent learning — that they can store information for later use without any process of reinforcement. One important follow-up question is: What sort of information is being stored? In particular, are the rats storing information about the spatial layout of the maze? Or are they simply "remembering" the sequences of movements (responses) that they made while wandering around the maze? And so, when the rats in the latent-learning experiments start running the maze successfully, are they simply repeating their earlier sequences of movements, or are they using their "knowledge" of how the different parts of the maze fit together?

Tolman and his students and collaborators designed many experiments during the 1930s and 1940s to try to decide between *place-learning* and

response-learning accounts of how rats learn to run a maze. Some of these experiments were reported in a famous article in 1946.

Cognitive Maps in Rats? Tolman et al., " Studies in Spatial Learning " (1946)

One experiment used a cross-maze with four end points (north, south, east, west), like that illustrated in Figure 1.3 . Rats were started at north and south on alternate trials. One group of rats was rewarded by food that was located at the same end point — east, for example. The relevant feature of the map for this group was that the same turning response would not invariably return them to the reward. To get from north to east, the rat needed to make a left-hand turn, whereas a right-hand turn was required to get from south to east. For the second group, the location of the food reward was shifted between east and west so that whether they started at north or south, the same turning response was required to obtain the reward. A rat in the second group starting from north would find the reward at east, whereas the same rat starting from south would find the reward at west. Whether the rat started at north or south, a left turn would always take it to the reward.

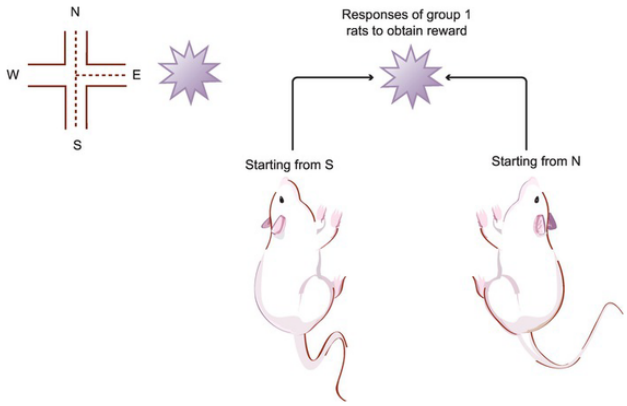


Figure 1.3 A cross-maze, as used by Tolman et al. (1946). The left-hand part of the figure illustrates the maze, with a star indicating the location of the food reward. The right-hand side illustrates how the group 1 rats had to make different sequences of movements in order to reach the reward, depending on where they started.

This simple experiment shows very clearly the distinction between place learning and response learning. Consider the first group of rats (those for which the food was always in the same place). They could only consistently obtain the reward if they represented it as being at a particular place and controlled their movements accordingly. Merely repeating the same response would only lead to success in half of the trials. For the second group, though, repeating the same turning response would invariably bring them to the reward, irrespective of the starting point.

Tolman found that the first group of rats learned to run the maze much more quickly than the second group. From this, he drew conclusions about the nature of animal learning in general, namely, that it was easier for animals to code spatial information in terms of places rather than in terms of particular sequences of movements .

Exercise 1.2 Explain in your own words why the experimental results seem to show that rats engage in place learning rather than response learning.

Tolman took his place-learning experiments as evidence that animals form high-level representations of how their environment is laid out — what he called *cognitive maps*. Tolman's cognitive maps were one of the first proposals for explaining behavior in terms of *representations* (stored information about the environment).

Representations are a fundamental explanatory tool in cognitive science. Cognitive scientists regularly explain particular cognitive achievements (such as the navigational achievements of rats in mazes) by modeling how the organism is using representations of the environment. Throughout this book, we will be looking at different models of how representations code information and how those representations are manipulated and transformed as the organism negotiates and engages with its environment.

Plans and Complex Behaviors: Lashley, "The Problem of Serial Order in Behavior" (1951)

At the same time as Tolman was casting doubt on standard behaviorist models of spatial navigation, psychologist and physiologist Karl Lashley was thinking more generally about the problem of explaining complex behavior.

Much human and animal behavior has a very complex structure, involving highly organized sequences of movements. Stimulus – response behaviorists have limited resources for thinking about these complex behaviors. They have to view them as linked sequences of responses — as a sort of chain, with each link determined by the link immediately preceding it. This is the basic idea behind response-learning models of how rats run mazes.

The standard behaviorist view is that rats learn to chain together a series of movements that leads to the reward. Tolman showed that this is not the right way to think about what happens when rats learn to run mazes. Lashley made the far more general point that this seems to be completely the wrong way to think about many complex behaviors.

Think of the complicated set of movements involved in uttering a sentence in English, for example. Or playing a game of tennis. In neither of these cases is what happens at a particular moment solely determined by what has just happened — or prompted by what is going on in the environment and influencing the organism. What happens at any given point in the sequence is often a function of what will happen later in the sequence, as well as of the overall goal of the behavior.

According to Lashley, many complex behaviors are products of prior planning and organization. They are organized hierarchically (rather than linearly). An overall plan (say, walking over to the table to pick up the glass) is implemented by simpler plans (the walking plan and the reaching plan), each of which can be broken down into simpler plans, and so on. Very little (if any) of this planning takes place at the conscious level.

Exercise 1.3 Give your own example of a hierarchically organized behavior.

Lashley ' s essay points toward two ideas that have proved very important for cognitive science. The first is the idea that much of what we do is under the control of planning and information-processing mechanisms that operate below the threshold of awareness. This is the *hypothesis of subconscious information processing*. Even though we are often conscious of our high-level plans and goals at the top of the hierarchy, we are typically not aware of the information processing translating those plans and goals into actions. So, for example, you might consciously form an intention to pick up a glass of water. But implementing that intention requires calculating very precisely the trajectory that your arm must take, as well as ensuring that your hand is open to the right degree to take hold of the glass. These calculations are carried out by information-processing systems operating far below the threshold of conscious awareness.

The second is the *hypothesis of task analysis*. This is the idea that we can understand a complex task (and the cognitive system performing it) by breaking it down into a hierarchy of more basic subtasks (and associated subsystems). This hypothesis has proven to be a powerful tool for understanding many different aspects of mind and cognition. We can think about a particular cognitive system (say, the memory system) as carrying out a particular task — the task of allowing an organism to exploit previously acquired information. We can think about that task as involving a number of simpler subtasks — say, the subtask of storing information and the subtask of retrieving information. Each of these subtasks can be carried

out by even more simple sub-subtasks. We might distinguish the sub-subtask of storing information for the long term from the sub-subtask of storing information for the short term. And so on down the hierarchy .

1.2 The Theory of Computation and the Idea of an Algorithm

At the same time as Tolman, Lashley, and others were putting pressure on some of the basic principles of behaviorism, the theoretical foundations for one highly influential approach to cognitive science (and indeed for our present-day world of omnipresent computers and constant flows of digital information) were laid in the 1930s, in what was at the time a rather obscure and little-visited corner of mathematics.

In 1936－1937, Alan Turing published an article in the *Proceedings of the London Mathematical Society* that introduced some of the basic ideas in the theory of computation. Computation is what computers do, and according to many cognitive scientists, it is what minds do. What Turing gave us was a theoretical model that many have thought to capture the essence of computation. Turing ′s model (the so-called *Turing machine*) is one of the most important and influential ideas in cognitive science, even though it initially seems to have little to do with the human mind.

Algorithms and Turing Machines: Turing, " On Computable Numbers, with an Application to the Decision Problem " (1936 – 1937)

Turing, together with a number of mathematicians working in the foundations of mathematics, was grappling with the problem (known as the *Halting Problem*) of finding a purely mechanical procedure for determining whether certain basic mathematical problems have a solution.

You can think about the Halting Problem in terms of computer programs. Many computer programs are not defined for every possible input. They will give a solution for some inputs, the ones for which they are defined. But for other inputs, the ones for which they are not defined, they will just endlessly loop, looking for a solution that isn ' t there. From the point of view of a computer programmer, it is really important to be able to tell whether or not the computer program is defined for a given input — in order to be able to tell whether the program is simply taking a very long time to get to the solution or whether it is in an endless loop.

This is what a solution to the Halting Problem would give — a way of telling, for a given computer program and a given input, whether the program is defined for that input. The solution has to work both ways. It has to give the answer " Yes " when the program is defined and " No " when the program is not defined.

Turing was looking for a purely mechanical solution to the Halting Problem. He was looking for something with the same basic features as the " recipes " that we all learn in high school for multiplying two numbers or performing long division. These recipes are mechanical because they do not involve any insight. The recipes can be clearly stated in a finite set of

instructions, and following the instructions correctly always gives the right answer, even if you don ′ t understand how or why.

Because the notion of a purely mechanical procedure is not itself a mathematical notion, the first step was to make it more precise. Turing did this by using the notion of an *algorithm.* An algorithm is a finite set of rules that are unambiguous and that can be applied systematically to an object or set of objects to transform it or them in definite and circumscribed ways. The instructions for programming a DVD recorder, for example, are intended to function algorithmically so that they can be followed blindly in a way that will transform the DVD recorder from being unprogrammed to being programmed to switch itself on and switch itself off at appropriate times. Of course, the instructions are not genuinely algorithmic because, as we all know, they are not idiot-proof.

Exercise 1.4 Think of an example of a genuine algorithm, perhaps from elementary arithmetic or perhaps from everyday life.

One of Turing ′ s great contributions was a bold hypothesis about how to define the notion of an algorithm within mathematics. Turing devised an incredibly simple kind of computing mechanism (now called, in his honor, a *Turing machine*). This is an idealized machine, not a real one. It consists of an infinitely long piece of tape divided into cells. The point of the tape being infinitely long is so that the machine will not have any storage limitations. A Turing machine is like a computer with an infinitely large hard disk. Turing did not think that a Turing machine would ever have to

deal with infinitely long strings of symbols. He just wanted it to be able to deal with arbitrarily long, but still finite, strings of symbols.

Each cell in the Turing tape is either blank or contains a single symbol. The tape runs through a machine head, with a single cell under the head at a given moment. This allows the head to read the symbol the cell contains. The machine head can also carry out a limited number of operations on the cell that it is currently scanning. It can do the following:

- Delete the symbol in the cell
- Write a new symbol in the cell
- Move the tape one cell to the left
- Move the tape one cell to the right

Any individual Turing machine has a set of instructions (its *machine table*). The machine can be in any one of a (finite) number of different states. The machine table determines what the Turing machine will do when it encounters a particular symbol in a particular cell, depending on which internal state it is in. Figure 1.4 is a schematic representation of a Turing machine.

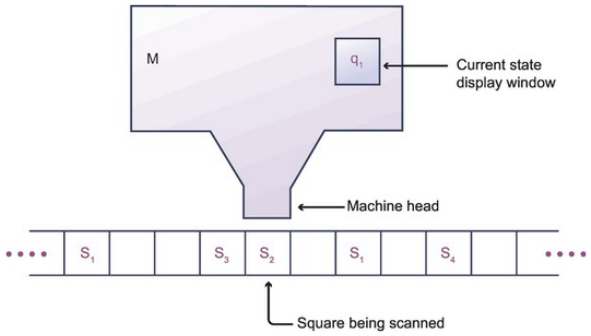


Figure 1.4 Schematic representation of a Turing machine.

(Adapted from Cutland 1980 .)

The beauty of a Turing machine is that its behavior is entirely determined by the machine table, its current state, and the symbol in the cell it is currently scanning. There is no ambiguity and no room for the machine to exercise " intuition " or " judgment. " It is, in fact, purely mechanical in exactly the way required for an algorithm.

Turing did not actually build a Turing machine. (It is difficult to build a machine with an infinitely long piece of tape!) But he showed how Turing machines could be specified mathematically. The machine table of a Turing machine can be represented as a sequence of numbers. This allowed him to prove mathematical results about Turing machines. In particular, it allowed him to prove that there is a special kind of Turing machine, a *universal Turing machine* , that can run any specialized Turing machine. The universal Turing machine can take as input a program specifying any given specialized Turing program. It is the theoretical precursor (with unlimited storage) of the modern-day general-purpose digital computer.

Turing ' s paper contained a subtle proof that the Halting Problem cannot be solved. It was also significant for articulating what we now call the *Church – Turing thesis* (in recognition of the contribution made by the logician Alonzo Church). According to the Church – Turing thesis, anything that can be done in mathematics by an algorithm can be done by a Turing machine. Turing machines are computers that can compute anything that can be algorithmically computed.

What Turing contributed to the early development of cognitive science (although at the time, his work was little known and even less appreciated) was a model of computation that looked as if it might be a clue to how information could be processed by the mind. As theorists moved closer to the idea that cognition involves processing information, it was an easy step

to think about information processing as an algorithmic process along the lines analyzed by Turing — a step that became even easier in the light of the huge advances that were made in designing and building digital computers (which, if the Church – Turing thesis is true, are essentially large and fast Turing machines) during and after the Second World War.

Exercise 1.5 Explain in your own words why the Church – Turing thesis entails that any computer running a program is simply a large and fast Turing machine.

1.3 Linguistics and the Formal Analysis of Language

The study of language played a fundamental role in the prehistory of cognitive science. On the one hand, language use is a paradigm of the sort of hierarchically organized complex behavior that Lashley was talking about. On the other hand, the emergence of transformational linguistics and the formal analysis of *syntax* (those aspects of language use that have to do with how words can be legitimately put together to form sentences) provided a very clear example of how to analyze, *in algorithmic terms* , the bodies of information that might underlie certain very basic cognitive abilities (such as the ability to speak and understand a language).

One crucial landmark was the publication in 1957 of *Syntactic Structures* by Noam Chomsky, a hugely important figure in the development of cognitive science. The transformational grammar proposed

by Chomsky (and subsequently much modified by Chomsky and others) reflects some of the basic ideas covered earlier in this chapter.

The Structure of Language: Chomsky ' s *Syntactic Structures* (1957)

Chomsky ' s book is widely held to be the first example of a linguist proposing an explanatory theory of *why* languages work the way they do (as opposed to simply describing and classifying *how* they work). Chomsky was interested not in mapping the differences between different languages and in describing their structure but, rather, in providing a theoretical account of why they have the structure that they do. Crucial to his approach is the distinction between the *deep structure* of a sentence (as given by what Chomsky calls a *phrase-structure grammar*) and its *surface structure* (the actual organization of words in a sentence, derived from the deep structure according to the principles of transformational grammar).

The deep structure, or phrase structure, of a sentence is simply how it is built up from basic constituents (syntactic categories) according to basic rules (phrase-structure rules). We only need a small number of basic categories to specify the phrase structure of a sentence. These are the familiar parts of speech that we all learn about in high school — nouns, verbs, adjectives, and so on. Any grammatical sentence (including those that nobody is ever likely to utter) is made up of these basic parts of speech combined according to basic phrase-structure rules (such as the rule that every sentence is composed of a verb phrase and a noun phrase).

In Figure 1.5 , we see how these basic categories can be used to give a phrase-structure tree of the sentence " John has hit the ball. " The phrase-

structure tree is easy to read, with a bit of practice. Basically, you start at the top with the most general characterization. As you work your way down the tree, the structure of the sentence becomes more finely articulated so that we see which words or combinations of words are doing which job.

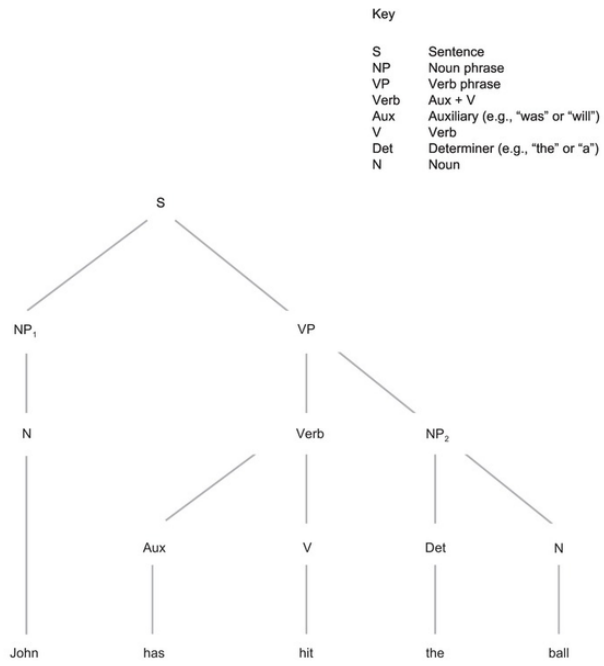


Figure 1.5 A sample phrase-structure tree for the sentence " John has hit the ball. " The abbreviations in the diagram are explained in the key.

Analyzing sentences in terms of their phrase structure is a powerful explanatory tool, but it has its limits. There are pairs of sentences that have very different phrase structures but are clearly very similar in meaning. Think of " John has hit the ball " and " The ball has been hit by John. " In most contexts, these sentences are equivalent and interchangeable, despite having very different phrase structures. Conversely, there are sentences with superficially similar phrase structures that are plainly unrelated. Think of " Susan is easy to please " and " Susan is eager to please. "

Exercise 1.6 Explain in your own words the difference between " Susan is easy to please " and " Susan is eager to please. " Why are their phrase structures different?

The basic aim of transformational grammar is to explain the connection between sentences of the first type and to explain the differences between sentences of the second type. This is done by giving principles that state the acceptable ways of transforming deep structures. This allows linguists to identify the transformational structure of a sentence in terms of its transformational history.

The transformational principles of transformational grammar are examples of *algorithms.* They specify a set of procedures that operate on a string of symbols to convert it into a different string of symbols. So, for example, our simple phrase-structure grammar might be extended to include an active – passive transformation rule that takes the following form (look at the key in Figure 1.5 for the translation of the symbols):

$$NP_1 + Aux + V + NP_2$$
$$\Rightarrow$$
$$NP_2 + Aux + been + V + by + NP_1$$

This transforms the string " John + has + hit + the + ball " into the string " The + ball + has + been + hit + by + John. " And it does so in a purely mechanical and algorithmic way.

Exercise 1.7 Write out an algorithm that carries out the active – passive transformation rule. Make sure that your algorithm instructs the person/machine following it on what to do at each step.

What ' s more, when we look at the structure of the passive sentence, " The ball has been hit by John, " we can see it as illustrating precisely the sort of hierarchical structure to which Lashley drew our attention. This is a characteristic of languages in general. They are hierarchically organized.

So, in thinking about how languages work, transformational grammar brings together two very fundamental ideas. The first idea is that a sophisticated, hierarchically organized cognitive ability, such as speaking and understanding a language, involves stored bodies of information (information about phrase structures and transformation rules). The second idea is that these bodies of information can be manipulated algorithmically .

1.4 Information-Processing Models in Psychology

In the late 1950s, the idea that the mind works by processing information began to take hold within psychology. A further influence was the emergence of information theory in applied mathematics. Rather unusually, this can be pinned down to a single event — the publication of an article entitled " A Mathematical Theory of Communication " by Claude E. Shannon in 1948. Shannon ' s paper showed how information can be

measured, and he provided precise mathematical tools for studying the transmission of information.

These tools (including the idea of a *bit* as a measure of information) proved very influential in psychology — and for cognitive science more generally. We can illustrate how information-processing models became established in psychology through two very famous publications from the 1950s.

The first, George Miller ' s article " The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information " (1956), used the basic concepts of information theory to identify crucial features of how the mind works. The second, Donald Broadbent ' s 1954 paper " The Role of Auditory Localization in Attention and Memory Span, " presented two influential experiments that were crucial in Broadbent ' s later putting forward, in his 1958 book *Perception and Communication* , one of the first information-processing models in psychology. The type of flowchart model that Broadbent proposed (as illustrated in Figure 1.6) has become a standard way for cognitive scientists to describe and explain different aspects of cognition.

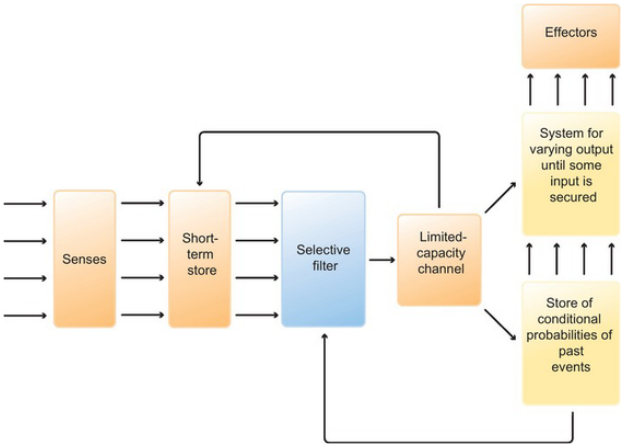


Figure 1.6 Donald Broadbent ' s (1958) model of selective attention.

(Redrawn by Derek Smith. From http://www.smithsrisca.demon.co.uk/PSYbroadbent1958.html .)

How Much Information Can We Handle? George Miller ' s " The Magical Number Seven, Plus or Minus Two " (1956)

One of the basic concepts of information theory is the concept of an information channel. An information channel is a *medium* that transmits information from a *sender* to a *receiver*. A telephone cable is an information channel. So is the radio frequency on which a television station broadcasts. Miller ' s insight was that perceptual systems, such as vision and hearing, are also information channels. This gave Miller and other psychologists a new set of tools for thinking about experiments on human perception.

Miller ' s article drew attention to a wide range of evidence suggesting that human subjects are really rather limited in the *absolute judgments* that they can make. An example of an absolute judgment is naming a color or identifying the pitch of a particular tone — as opposed to relative judgments, such as identifying which of two colors is the darker one or which of two tones is higher in pitch.

In one experiment reported by Miller, subjects are asked to assign numbers to the pitches of particular tones and then presented with sequences of tones and asked to identify them in terms of the assigned numbers. So, for example, if you assigned " 1 " to middle C, " 2 " to the first E above middle C, and " 3 " to the first F# and then heard the sequence E-C-C-F#-E, the correct response would be 2-1-1-3-2.

When the sequence is only one or two tones long, subjects never make mistakes. But performance falls off drastically when the sequence is six or more tones long. A similar phenomenon occurs when we switch from hearing to vision and ask subjects to judge the size of squares or the length of a line. Here, too, there seems to be an upper bound on the number of distinct items that can be processed simultaneously.

Putting these (and many other) experimental results into the context of information theory led Miller to propose that our sensory systems are all information channels with roughly the same *channel capacity* (where this is the amount of information an information channel can reliably transmit). The perceivers ' capacity to make absolute judgments is an index of the channel capacity of the information channel that they are using.

In effect, Miller suggested, human perceptual systems are information channels with built-in limits. They can only process around seven items at the same time (or to put it in the language of information theory, their channel capacity is around 3 bits; because each bit allows the system to discriminate 2 pieces of information, n bits of information will allow the system to discriminate 2^n pieces of information, and 7 is just under 2^3).

At the same time as identifying these limits, Miller identified ways of working around them. One way of increasing the channel capacity is to *chunk* information. We can relabel sequences of numbers with single

numbers. A good example (discussed by Miller) comes when we use decimal notation to relabel numbers in binary notation. We can pick out the same number in two different ways — with the binary expression 1100100, for example, or with the decimal expression 100. If we use binary notation, then we are at the limits of our visual channel capacity. If we use decimal notation, then we are well within those limits. As Miller pointed out, natural language is the ultimate chunking tool.

Exercise 1.8 Think of an informal experiment that you can do to illustrate the significance of chunking information.

The Flow of Information: Donald Broadbent ' s " The Role of Auditory Localization in Attention and Memory Span " (1954) and *Perception and Communication* (1958)

The experiments reported and analyzed by Miller made plausible the idea that the senses are information channels with limited capacity. The obvious next step was to think about how those information channels actually work. One of the first models of how sensory information is processed was developed by the British psychologist Donald Broadbent in his 1958 book *Perception and Communication.* As with Miller, the impetus came from experiments in the branch of psychology known as *psychophysics* . This is the branch of psychology that studies how subjects perceive and discriminate physical stimuli.

We can appreciate what is going on by thinking about the so-called *cocktail party phenomenon* . When at a cocktail party, or any other social gathering, we can often hear many ongoing and unrelated conversations.

Somehow, we manage to focus only on the one we want to listen to. Auditory attention is selective. How do we manage this? How do we screen out all the unwanted sentences that we hear?

Broadbent used *dichotic listening experiments* , in which subjects are presented with different information in each ear. The experiments reported in his paper " The Role of Auditory Localization in Attention and Memory Span " (Broadbent 1954) involved presenting subjects with a string of three different stimuli (letters or digits) in one ear while simultaneously presenting them with a different string in the other ear. The subjects were asked to report the stimuli in any order. Broadbent found that they performed best when they reported the stimuli ear by ear — that is, by reporting all three presented to the left ear first, followed by the three presented to the right ear. This finding and others were explained by the model that he subsequently developed.

The basic features of the model are illustrated in Figure 1.6 . Information comes through the senses and passes through a short-term store before passing through a selective filter. The selective filter screens out a large portion of the incoming information, selecting some of it for further processing. This is what allows us to attend selectively to only a portion of what is going on around us in the cocktail party. Only information that makes it through the selective filter is semantically interpreted, for example. Although people at cocktail parties can hear many different conversations at the same time, many experiments have shown that they have little idea of what is said in the conversations that they are not attending to. They hear the words but do not extract their meaning.

Broadbent interpreted the dichotic listening experiments as showing that we can only attend to a single information channel at a time (assuming

that each ear is a separate information channel) — and that the selection between information channels is based purely on the physical characteristics of the signal. The selection might be based on the physical location of the sound (whether it comes from the left ear or the right ear, for example) or on whether it is a man ' s voice or a woman ' s voice.

The selective filter does not work by magic. As the diagram in Figure 1.6 shows, the selective filter is " programmed " by another system that stores information about the relative likelihoods of different events. We are assuming that the system is pursuing a goal. What is programming the selective filter is information about the sorts of things that have led to that goal being satisfied in the past. Information that makes it through the selective filter goes into what Broadbent calls the *limited-capacity channel* . Information that is filtered out is assumed to decay quickly. From the limited-capacity channel, information can go either into the long-term store or on to further processing and eventually into action, or it can be recycled back into the short-term store (to preserve it if it is in danger of being lost).

Exercise 1.9 Give an example in your own words of selective attention in action. Incorporate as many different aspects of Broadbent ' s model as possible .

1.5 Connections and Points of Contact

The most basic concept that has run through the chapter is the concept of information.

- Tolman ' s latent-learning experiments seemed to show that animals can pick up information without any reinforcement taking place.
- Chomsky ' s *Syntactic Structures* developed the idea that speaking and understanding natural languages depend on information about sentence structure — about the basic rules that govern the surface structure of sentences and about the basic transformation principles that underlie the deep structure of sentences.
- Miller and Broadbent explored perceptual systems as information channels and used the concepts of information theory to explore their basic structure and limits.

Hand in hand with the concept of information goes the concept of representation. Information is everywhere, but in order to use it, organisms need to represent it. Representations will turn out to be the basic currency of cognitive science, and we have seen several different examples in this chapter:

- Tolman ' s place-learning experiments introduced the idea that organisms have cognitive maps representing the spatial layout of the environment.
- Turing machines represent the instructions for implementing particular algorithms in their machine tables.
- Chomsky suggested that important elements of linguistic understanding are represented as phrase-structure rules and transformational rules.

- Miller showed how representing information in different ways (in terms of different types of chunking, for example) can affect how much information we are able to store in memory.

But what about the mechanics of information processing? How does it actually work?

- Turing machines offer a purely mechanical way of solving problems and processing information. In one sense, Turing machines are completely unintelligent. They blindly follow very simple instructions. And yet, if the Church－Turing thesis is warranted, they can compute anything that can be algorithmically computed. And so, in another sense, it would be difficult to be more intelligent than a Turing machine.
- In transformational linguistics, the basic principles that transform sentences (that take a sentence from its active to its passive form, for example) can be thought of as mechanical procedures that can in principle be carried out by a suitably programmed Turing machine.
- Broadbent＇s model of selective attention is built around the idea that information processing is done by dedicated and specialized systems. The complex information-processing task of making sense of the vast amounts of information picked up by the hearing system is broken down into a number of simpler tasks (such as the task of selecting a single information channel or the task of working out what sentences mean). Each information-processing task is performed by a dedicated system, such as the selective filter or the semantic-processing system.

- Also emerging from Broadbent ′ s model is the idea that we *can understand how a cognitive system as a whole works by understanding how information flows through the system.* Broadbent offered a flowchart showing the different stages that information goes through as it is processed by the system, and many psychologists and cognitive scientists subsequently took this type of information-processing flowchart to be a paradigm of how to explain cognitive abilities.

Summary

This chapter has surveyed four of the most important precursors of what subsequently became known as *cognitive science* . Cognitive science emerged when experimentalists and theoreticians began to see connections between developments in disciplines as diverse as experimental psychology, theoretical linguistics, and mathematical logic. These connections converge on the idea that cognition is a form of information processing and hence that we can understand how the mind works and how organisms negotiate the world around them by understanding how information about the environment is represented, transformed, and exploited.

Checklists

Important Developments Leading Up to the Emergence of Cognitive Science

1. The reaction against behaviorism in psychology

2. Theoretical models of computation from mathematical logic

3. Systematic analysis of the structure of natural language in linguistics

4. The development of information-processing models in psychology

Central Themes of the Chapter

1. Even very basic types of behavior (such as the behavior of rats in mazes) seem to involve storing and processing information about the environment.

2. Information relevant to cognition can take many forms — from information about the environment to information about how sentences can be constructed and transformed.

3. Perceptual systems can be viewed as information channels, and we can study both (a) the very general properties of those channels (e.g., their channel capacity) and (b) the way in which information flows through those channels.

4. Mathematical logic and the theory of computation show us how information processing can be mechanical and algorithmic.

5. Much of the information processing that goes on in the mind takes place below the threshold of awareness.

Further Reading

The story of how cognitive science emerged is told in Gardner ' s *The Mind ' s New Science* (1985). Flanagan ' s *The Science of the Mind* (1991) goes

further back into the prehistory of cognitive science and psychology, as do the papers in Brook 2007 . Margaret Boden ' s two-volume *Mind as Machine: A History of Cognitive Science* (2006) is detailed, placing most emphasis on computer science and artificial intelligence. Abrahamsen and Bechtel ' s chapter in Frankish and Ramsey 2012 provides a concise summary of the history of cognitive science. Going back still further, histories of psychology, such as Hergenhahn and Henley 2013 , typically start with the ancient Greeks.

The basic principles of classical and operant conditioning are covered in standard textbooks on psychology, such as Gazzaniga et al. 2011 , Plotnik and Kouyoumdjian 2010 , and Kalat 2010 . For a survey of contemporary research on both types of conditioning, see McSweeney and Murphy 2014 . Watson ' s article " Psychology as the Behaviorist Views It " is a classic behaviorist manifesto (Watson 1913). It can be found in the online resources. Tolman ' s article " Cognitive Maps in Rats and Men " (1948) gives an accessible introduction to many of his experiments and is also in the online resources. Gallistel 1990 is a very detailed and sophisticated presentation of a computational approach to animal learning.

Turing ' s paper on undecidable propositions (Turing 1936 – 1937) will defeat all but graduate students in mathematical logic (but see Petzold 2008 for a book-length explanation aimed at the general reader). His paper " Computing Machinery and Intelligence " (Turing 1950) is a much more accessible introduction to his thoughts about computers. There are several versions online, the best of which are included in the online resources. Hodges 2014 is a new edition of the classic biography of Turing, which inspired the film *The Imitation Game.* Martin Davis has written two popular books on the early history of computers: Davis 2000 and 2001 . Copeland

1993 gives a more technical but still accessible account of Turing machines and the Church – Turing thesis. Millican and Clark 1996 is a collection of papers on Turing ′ s legacy. A more general article illustrating the algorithmic nature of information processing is Schyns et al. 2008 .

At more or less the same time as Turing was working on the mathematical theory of computation, the neurophysiologist Warren McCulloch and logician Walter Pitts were collaborating on applying rather similar ideas about computation directly to the brain. Their paper " A Logical Calculus of the Ideas Immanent in Nervous Activity " (McCulloch and Pitts 1943) was influential at the time, particularly in the early development of digital computers, but is rarely read now. It is reprinted in Cummins and Cummins 2000 . An accessible survey of their basic ideas can be found in Anderson 2003 . See also Chapter 2 of Arbib 1987 and Piccinini 2004 , as well as Schlatter and Aizawa 2008 .

Most people find Chomsky ′ s (1957) *Syntactic Structures* pretty hard-going. Linguistics tends to be technical, but Chomsky ′ s article " Linguistics and Philosophy, " reprinted in Cummins and Cummins 2000 , contains a fairly informal introduction to the basic distinction between surface structure and deep structure. Chapter 2 of Newmeyer 1986 is a good and accessible introduction to the Chomskyan revolution. More details can be found in standard textbooks, such as Cook and Newson 2007 , Isac and Reiss 2013 , and O ′ Grady et al. 2010 . Chomsky ′ s rather harsh review of B. F. Skinner ′ s book *Verbal Behavior* (Chomsky 1959) is often described as instrumental in the demise of radical behaviorism — and hence in bringing about the so-called *cognitive revolution* . The review is reprinted in many places and can be found in the online resources. Pinker 1994 presents a broadly Chomskyan perspective on language for a general audience.

The cocktail party phenomenon was first introduced in Cherry 1953 . For a concise summary, see McDermott 2009 . Miller ' s 1956 article is widely available and is included in the online resources. Broadbent ' s model of selective attention was the first in a long line of models. These are reviewed in standard textbooks, such as Chapter 5 of Gleitman et al. 2010 . Christopher Mole ' s chapter on attention in Margolis et al. 2012 summarizes Broadbent ' s influence as well as recent departures from Broadbent. Driver 2001 is an article-length survey of theories of selective attention in the twentieth century.

Chapter 2

The Discipline Matures: Three Milestones

◈

Overview

2.1 Language and Micro-Worlds

Natural Language Processing: Winograd, *Understanding Natural Language* (1972)

SHRDLU in Action

2.2 How Do Mental Images Represent?

Mental Rotation: Shepard and Metzler, " Mental Rotation of Three-Dimensional Objects " (1971)

Information Processing in Mental Imagery

2.3 An Interdisciplinary Model of Vision

Levels of Explanation: Marr ' s *Vision* (1982)

Applying Top-Down Analysis to the Visual System

Overview

Chapter 1 explored the prehistory of cognitive science in the first half of the twentieth century. We turn now to three milestones in the development of cognitive science proper. In each of them, we see the theoretical ideas canvassed in the previous section being combined and applied to understanding specific cognitive systems and cognitive abilities.

Section 2.1 looks at a powerful and influential computer model of what it is to understand a natural language. Terry Winograd ' s computer model SHRDLU illustrates how grammatical rules might be represented *algorithmically* in a cognitive system and integrated with other types of information about the environment.

The idea that the digital computer is the most promising model for understanding the mind was at the forefront of cognitive science in the 1960s and 1970s. But even in the 1970s, it was under pressure. Section 2.2 looks at an important debate about mental imagery. Some very influential experiments in cognitive psychology seemed to show that imagistic reasoning involves representations very different from those in a digital computer.

The third section introduces what many cognitive scientists still consider to be cognitive science ' s greatest single achievement: the theory of early visual processing developed by David Marr. Marr ' s theory of vision was highly interdisciplinary, drawing on mathematics, cognitive psychology, neuroscience, and the clinical study of patients with brain damage, and it was built on a hierarchy of different levels for studying cognition that was, for a long time, taken to define the method of cognitive science.

2.1 Language and Micro-Worlds

The human ability to speak and understand natural language is one of our most sophisticated cognitive achievements. We share many types of cognitive abilities with nonlinguistic animals. There are significant continuities between human perceptual systems and those of the higher primates (such as chimpanzees and macaque monkeys), which is why much of what we know about the neural structure of the human perceptual system is actually derived from experiments on monkeys. (More on this in Chapters 3 and 9 .) And there is powerful evidence that prelinguistic infants are capable of representing and reasoning about their physical and social environments in comparatively sophisticated ways. (See Chapter 11 for more details.)

Nonetheless, just as in human development (*ontogeny*) there is a cognitive explosion that runs more or less in parallel with the acquisition of language, much of what distinguishes humans from other animals is intimately bound up with our linguistic abilities. Language is far more than a tool for communication. It is a tool for thinking. Without language, there would be no science and no mathematics. Language allows us to engage in incredibly sophisticated types of coordinated behavior. It underpins our political and social structures. In many ways, *Homo linguisticus* would be a more accurate name than *Homo sapiens*.

So, if cognitive scientists want to understand the human mind, then they have to confront the fundamental challenge posed by our understanding of natural language. As we saw in the last chapter, Chomsky ' s transformational linguistics was very important in setting the stage for the cognitive revolution. But Chomsky has relatively little to say about how

linguistic understanding actually takes place. It is one thing to describe the abstract structure of human language and quite another to explain how human beings can master that abstract structure. Winograd ′ s work represents an early attempt to fill that gap.

Natural Language Processing: Winograd, *Understanding Natural Language* (1972)

The early days of *artificial intelligence* (AI) saw a number of attempts to write computer programs that could engage in some very elementary forms of conversational exchanges, but none of these programs was capable of anything that really resembled linguistic understanding.

Programs such as ELIZA (written by Joseph Weizenbaum in 1966) aimed to simulate human conversation. The basic idea behind ELIZA (which, depending upon whom one asks, was either based on or intended to parody typical conversational exchanges between psychotherapists and their patients) was to create the illusion of conversation by rephrasing statements as questions and by programming the computer to give certain fixed responses where this is not possible. A sample " conversation " is given in Box 2.1 .

BOX 2.1 A Conversation with ELIZA

In the following exchange, ELIZA ′ s responses are in uppercase.

Men are all alike.

IN WHAT WAY?

They are always bugging us about something or other.

CAN YOU THINK OF A SPECIFIC EXAMPLE?

Well, my boyfriend made me come here.

YOUR BOYFRIEND MADE YOU COME HERE?

He says I ′ m depressed much of the time.

I ′ M SORRY TO HEAR THAT YOU ARE DEPRESSED.

It ′ s true. I am unhappy.

DO YOU THINK COMING HERE WILL HELP YOU NOT TO BE UNHAPPY?

I need some help. That much seems certain.

WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP?

Perhaps I could learn to get along with my mother.

TELL ME MORE ABOUT YOUR FAMILY.

(From Boden 1977 : 106 – 107.)

Although ELIZA is said to have fooled a number of people into thinking that it was a human (including the unknowing participant in the conversation recorded in the box), nobody has ever suggested that it displays anything like a genuine understanding of language. ELIZA does not, in any sense, analyze the syntactic structure or the meaning of the sentences that it encounters. It is simply programmed to respond to certain cues by making one of a small set of responses. Nor could ELIZA use the conversations in which it engaged to report on or navigate its environment. *Chatbot* programs such as ELIZA are interesting for many reasons — but not as serious models of how we understand and use language.

Exercise 2.1 Explain in your own words what you think we can learn from programs such as ELIZA. Is it important that a person might be fooled by ELIZA into thinking that they were communicating with another human being?

Terry Winograd ' s program SHRDLU, initially presented in his 1970 doctoral dissertation at the Massachusetts Institute of Technology (MIT), was one of the first attempts to write a program that was not just trying to simulate conversation but that was capable of using language to report on its environment, to plan actions, and to reason about the implications of what is being said to it.

One of the distinctive features of SHRDLU is that it is programmed to deal with a very limited *micro-world* (as opposed to being a general-purpose language program, which is what ELIZA and other chatterbot programs are, in their very limited ways). The SHRDLU micro-world is very simple. It

consists simply of a number of colored blocks, colored pyramids, and a box, all located on a tabletop, as illustrated in Figure 2.1 . (Note that the micro-world is completely virtual. Everything takes place on a computer screen.)

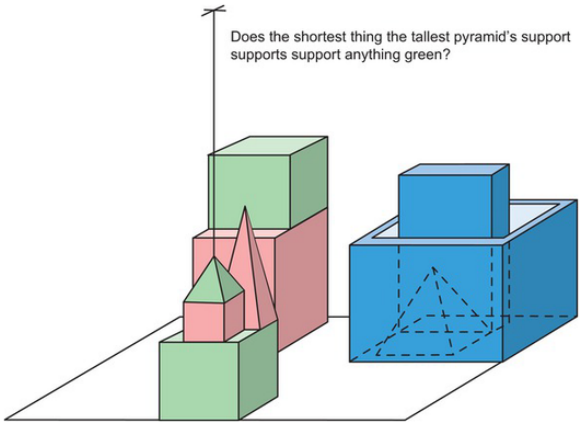


Figure 2.1 A question for SHRDLU about its virtual micro-world.

(Adapted from Winograd 1972 .)

SHRDLU can carry out various actions in the micro-world using a (virtual) robot arm. It can pick up the blocks and pyramids, move them around, and put them in the box. Corresponding to the simplicity of the micro-world, SHRDLU ' s language is relatively simple. It only has the tools to talk about what is going on in the micro-world.

SHRDLU was very important in the development of cognitive science, for three main reasons. First, it gave a powerful illustration of how abstract rules and principles such as those in the sort of grammar that we might find in theoretical linguistics could be practically implemented. SHRDLU provided a model of how a speaker ' s linguistic knowledge could be *represented* by a cognitive system and *integrated* with other, more general forms of knowledge about the environment.

Second, SHRDLU illustrates the general approach of trying to understand and model cognitive systems by breaking them down into distinct components, each carrying out a specific information-processing task. These distinct components are not completely self-contained. They collaborate in solving information-processing problems because the programs for each processing system allow it to consult other processing systems at particular moments in the computation.

Finally, SHRDLU is based on the fundamental assumption that understanding language is an *algorithmic* process. In Winograd ′ s own words, " All language use can be thought of as a way of activating procedures within the hearer " (1973: 104). Each component system is made up of a vast number of procedures that work algorithmically to solve very specific problems. The system as a whole works because of how these procedures are linked up and embedded within each other.

SHRDLU in Action

As is often the case in so-called *classical cognitive science* , the best way to understand what is going on in SHRDLU is to work from the top down — to start by looking at the general overall structure and then drill down into the details. SHRDLU consists of twelve different systems, which Winograd divides into three groups, each carrying out a specific job. These are exactly the jobs that one would expect any language-processing system to perform.

1. *Syntactic analysis:* SHRDLU needs to *parse* the grammatical structure of the sentences that it encounters, identifying which units in the sentence are performing which linguistic function — for example,

which linguistic units are functioning as nouns (picking out objects) and which as verbs (characterizing events and processes).

2. *Semantic analysis:* The system also needs to assign meanings to the individual words in a way that reveals what the sentence is stating (if it is a statement) or requesting (if it is a request). This takes us from *syntax* to *semantics* .

3. *Integrating the information acquired with the information the system already possesses:* The system has to be able to explore the implications of what it has just learned for the information it already has or to call on information it already has in order to obey some command, fulfill a request, or answer a question.

Winograd identifies distinct components for each of these jobs — the *syntactic system* , the *semantic system* , and the *cognitive-deductive system.* But they do not operate in strict sequence. The syntactic system does not produce a syntactic analysis, then hand that syntactic analysis over to the semantic system so that it can plug meanings into the abstract syntactic structure before passing the result on to the cognitive-deductive system. In SHRDLU, all three systems operate concurrently and can call upon each other at specific points.

This is possible because although all three systems store and deploy different forms of knowledge, they are all represented in a similar way. They are all represented in terms of *procedures.*

The best way to understand what procedures are is to look at some examples. One very fundamental " decision " that the syntactic system has to make is whether its input is a sentence or not. Assume that we are dealing with a very simple language that only contains words in the following

syntactic categories: noun (e.g., " block " or " table "), intransitive verb (e.g., " ___ is standing up "), transitive verb (e.g., " ___ is supporting ___ "), determiner (e.g., " the " or " a "). Figure 2.2 presents a simple procedure for answering this question. The SENTENCE program exploits the fact that every grammatical sentence must contain a noun phrase (NP) and a verb phrase (VP). It tests for the presence of an NP, tests for the presence of a VP, and then checks that there is no extra " junk " in the sentence.

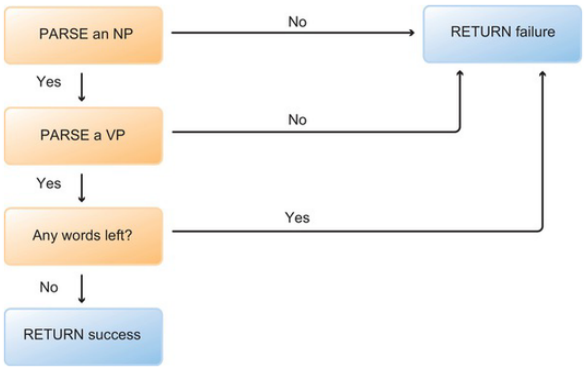


Figure 2.2 An algorithm for determining whether a given input is a sentence or not.

(Adapted from Winograd 1972 .)

Of course, in order to apply this procedure, the syntactic system needs procedures for testing for the presence of NPs and VPs. This can be done in much the same way — by checking, in an algorithmic manner, whether the relevant syntactic units are present. Figure 2.3 gives two procedures that will work in our simple language.

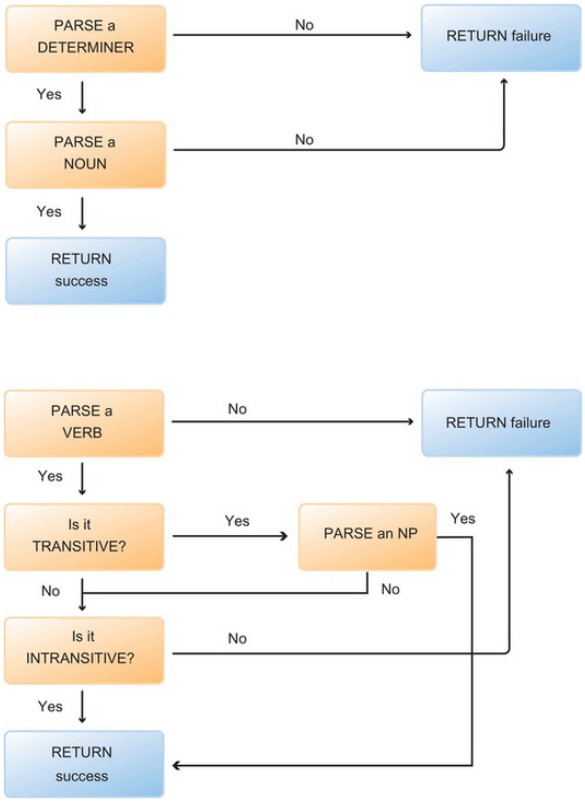


Figure 2.3 Algorithms for identifying noun phrases and verb phrases.

(Adapted from Winograd 1973 .)

Moving to semantic analysis, SHRDLU represents the meanings of words by means of comparable procedures. But instead of picking out syntactic categories, these procedures involve information about the micro-world and actions that the system can perform in the micro-world.

One of the words in SHRDLU ' s vocabulary is *CLEARTOP* . We can say that something (e.g., a block) is CLEARTOP when it does not have anything on it. CLEARTOP can also function as a command (as the command to remove anything resting on the block). CLEARTOP is represented in SHRDLU by the very simple procedure presented in Figure 2.4 .

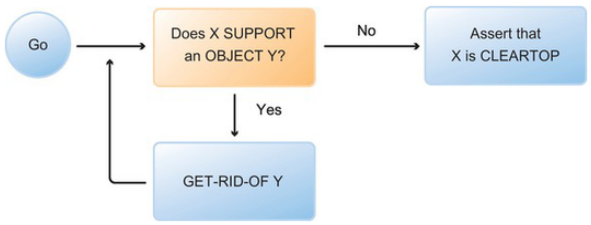


Figure 2.4 Procedure for applying the command CLEARTOP.

(Adapted from Winograd 1972 .)

This definition of CLEARTOP exploits other " concepts, " such as SUPPORT and GET RID OF. Each of these other concepts has its own procedure, which may well call upon the CLEARTOP procedure.

To see how these procedures work to allow SHRDLU to follow instructions in the micro-world and answer questions about it, we can look at the first few exchanges in a sample dialog described by Winograd in his 1973 paper. SHRDLU ' s contributions are in capital letters, whereas the sentences in italics were typed by a person. Normal type indicates Winograd ' s comments or author commentary.

1. *Pick up a big red block.*

OK.

The system answers " OK " when it carries out a command. In order to pick up the red block, it had to clear it off by finding a space for the green one and moving it away. The result is shown in Figure 2.5 .

2. *Grasp the pyramid.*

I DON ' T UNDERSTAND WHICH PYRAMID YOU MEAN.

There are three pyramids in the scene, and the program knows that phrases beginning with " the " are intended to refer to a specific object the speaker has in mind. So SHRDLU does nothing.

3. *Find a block that is taller than the one you are holding and put it in the box.*

The situation is depicted in Figure 2.5 . The system has to figure out what is meant by " one " and " it, " as well as use deduction to decide which block it is holding.

BY " IT, " I ASSUME YOU MEAN THE BLOCK THAT IS TALLER THAN THE ONE I AM HOLDING.

There are two possible referents for " it " : the one selected and " the one you are holding. "

OK.

The hand has to set down the red block to pick up the taller blue one. We are now at Figure 2.6 .

4. *What does the box contain?*

THE BLUE PYRAMID AND THE BLUE BLOCK.

5. *What is the pyramid supported by?*

THE BOX.

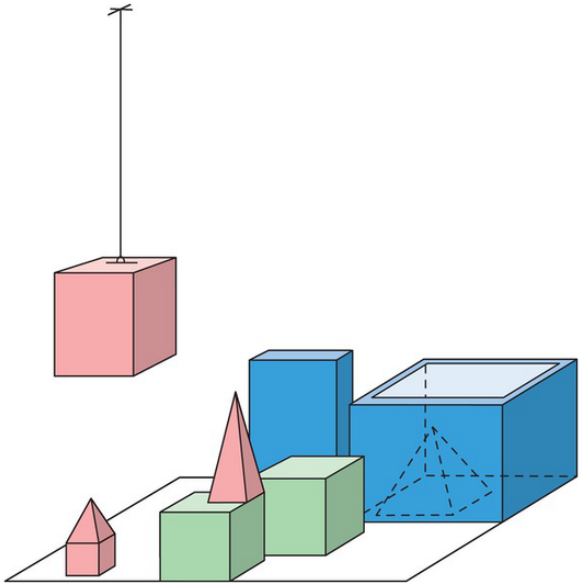


Figure 2.5 SHRDLU acting on the initial command to pick up a big red block. See the dialog in the text for what led up to this.

(Adapted from Winograd 1972 : 8.)

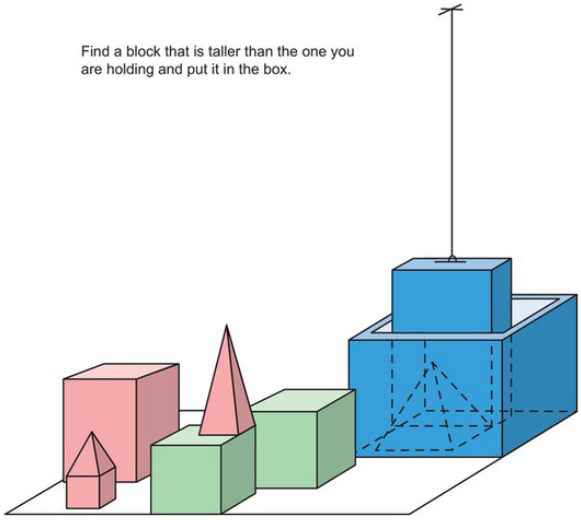


Figure 2.6 SHRDLU completing instruction 3 in the dialog: “ Find a block that is taller than the one you are holding and put it in the box. ”

(Adapted from Winograd 1972 : Figure 3.)

Earlier, the system responded to " the pyramid " by saying, " I DON ' T UNDERSTAND WHICH PYRAMID YOU MEAN. " Now SHRDLU has identified the pyramid that was meant — the pyramid just mentioned in the response to the previous question.

SHRDLU has its limitations. No one has ever claimed that it provides an accurate model of even a small segment of human language use. Winograd himself was quick to acknowledge that SHRDLU does not do justice to how conversations actually work in real life. Conversations are social interactions, not simply sequences of unconnected questions and answers. They involve agents trying to make sense of the world and of each other simultaneously.

Every level of linguistic understanding involves assumptions and guesses about what the other partner in the conversation is trying to communicate. It also involves assumptions and guesses about what they are trying to achieve. These are not always the same. In making and assessing those assumptions and guesses, we use all sorts of heuristics and principles. We tend to assume, for example, that people generally tell the truth, that they don ' t say things that are pointless and uninformative, and that what they say reflects what they are doing more generally. This is all part of what linguists call the *pragmatics* of conversation. But there is nothing comparable in SHRDLU ' s programming.

Exercise 2.2 Can you identify other ways in which SHRDLU falls short of modeling human understanding of natural language?

But nonetheless, SHRDLU was genuinely pathbreaking because it illustrates how a form of linguistic understanding can result from the interaction of many independently specifiable cognitive processes. Each process does a particular job — the job of identifying noun phrases, for example. These cognitive processes are themselves understood algorithmically (although this is not something that Winograd himself stresses). They involve processing inputs according to rules. Winograd ' s procedures are sets of instructions that can be followed mechanically, just as in the classical model of computation (see Section 1.2).

2.2 How Do Mental Images Represent?

One way to understand a complex cognitive ability is to build a machine that has that ability (or at least some primitive form of it). The program is, in effect, a model of the ability. This is why the history of cognitive science has been closely bound up with the history of AI.

AI can function as a form of experimentation. Particular ideas about how the mind works are written into programs, and then we " test " those ideas by seeing how well the programs work. But AI is not the only way of developing and testing hypotheses open to cognitive scientists. Cognitive scientists have also learned much from the more direct forms of experimentation carried out by cognitive psychologists.

As we saw in the previous chapter , the emergence of cognitive psychology as a serious alternative to behaviorism in psychology was one of the key elements in the emergence of cognitive science. A good example of how cognitive psychology can serve both as an inspiration and as a tool for

cognitive science came with what has come to be known as the *imagery debate* . This began in the early 1970s, inspired by a thought-provoking set of experiments on mental rotation carried out by psychologist Roger Shepard in collaboration with Jacqueline Metzler, Lynn Cooper, and other scientists.

The initial experiments are rightly recognized as classics of cognitive psychology. From the perspective of cognitive science, however, what is most interesting about them is the theorizing to which they gave rise about the format in which information is stored and the way in which it is processed. This was one of the first occasions when cognitive scientists seriously came to grips with the nature and format of mental representation — a theme that has dominated cognitive science ever since.

Mental Rotation: Shepard and Metzler, " Mental Rotation of Three-Dimensional Objects " (1971)

The original mental rotation experiments are easy to describe. Subjects were presented with drawings of pairs of three-dimensional figures. Figure 2.7 contains examples of these pairs.

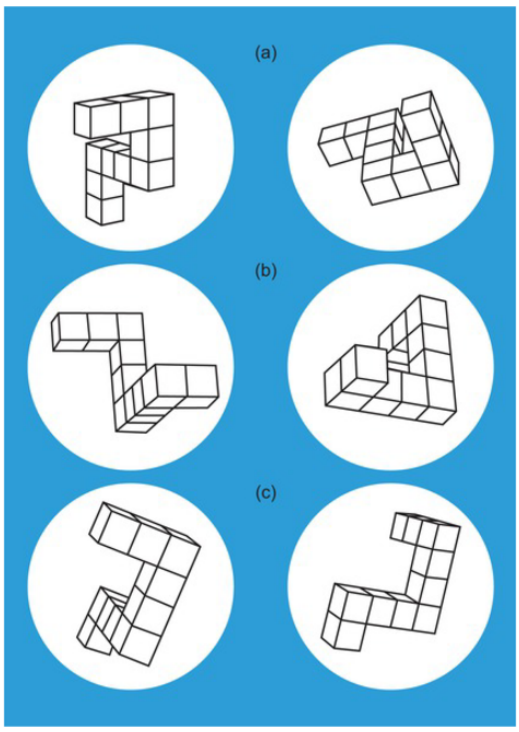


Figure 2.7 Examples of the three-dimensional figures used in Shepard and Metzler ′ s (1971) studies of mental rotation. Subjects were asked to identify which pairs depicted the same figure at different degrees of rotation.

(Adapted from Shepard and Metzler 1971 .)

Each figure is asymmetric and resembles its partner. In two cases, the figures resemble each other because they are, in fact, the same figure at different degrees of rotation. In a third case, the figures are different. The subjects were asked to identify, as quickly as possible, pairs of drawings where the second figure is the same as the first but rotated to a different angle. (You can do this experiment for yourself. Several versions of the Shepard – Metzler paradigm can be carried out online. See the Further Reading for an example. Putting " mental rotation " into a search engine will find others.)

Exercise 2.3 Which pair is the odd one out? In the pair with two distinct figures, how are those figures related to each other?

Shepard and Metzler found that there is a direct, linear relationship between the length of time that subjects took to solve the problem and the degree of rotation between the two figures (see Figure 2.8). The larger the angle of rotation, the longer subjects took to work out correctly that the two drawings depicted the same figure. And the length of time increased in direct proportion to the degree of rotation. These findings have proven to be very robust, with many follow-up experiments finding comparable effects. Much more controversial is how to interpret what is going on.

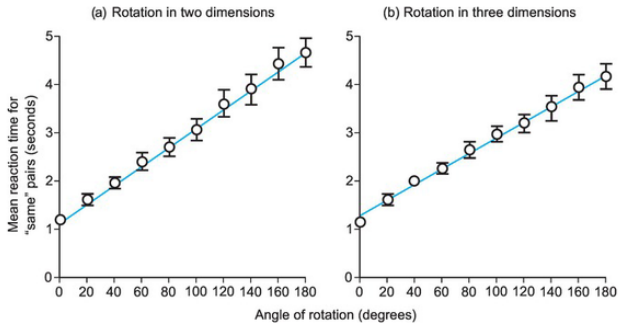


Figure 2.8 Results of Shepard and Metzler ′ s (1971) studies of mental rotation. (a) Mean reaction time for shape rotation in two dimensions. (b) Mean reaction time for shape rotation in three dimensions.

The subjects in the original experiment were not asked to solve the problem in any particular way. They were simply asked to pull one lever if the two pictures represented the same figure and another lever if the pictures represented different figures. The explanation that comes quickest to mind, though, is that the subjects solved the problem by mentally rotating one

figure to see whether or not it could be mapped onto the other. This would certainly provide a neat explanation of the findings. And this is indeed how Shepard, Metzler, and many others did interpret them (not least because that is what many of the subjects described themselves as doing). This interpretation of the experiments raises some fundamental questions about the format in which information is encoded and manipulated in tasks of this type.

Exercise 2.4 Present Shepard and Metzler ′ s conclusion in your own words. Explain their reasoning. What sort of assumptions does it rest on?

Suppose that we take the subject ′ s report of what they are doing in the experiments at face value. Suppose, that is, that we think of the subjects as rotating mental images in their " mind ′ s eye. " It seems, on the face of it, that this is really just an application of a skill that we use all the time — the skill of transforming mental images in order to calculate, for example, whether one ′ s car will fit into a tight parking space or where a tennis ball will land.

But appealing to rotation in my " mind ′ s eye " does not really solve the problem. Most cognitive scientists think that our conscious experience of rotating a mental image is the result of unconscious information processing. Information about the shape is derived from perception and then transformed in various ways that enable the subject to determine whether the two drawings are indeed drawings of the same shape. So the important question is: How is that information represented, and how is it transformed?

Information Processing in Mental Imagery

The standard way of thinking about the mind as an information processor takes the digital computer as a model. (This was almost unchallenged in the early 1970s and remains a popular view now, although we currently have a much clearer sense of some alternative ways of thinking about information processing.) Digital computers store and manipulate information in a fixed format. Essentially, all forms of information in a digital computer are represented using the binary numerals 0 and 1.

Each binary digit carries a single unit of information (a *bit*). Within the computer, these units of information are grouped into words — a *byte* , for example, is an eight-bit word that can carry 256 units of information. This way of carrying information in discrete quantities is often called *digital information storage* .

One feature of digitally encoded information is that the length of time it takes to process a piece of information is typically a function only of the quantity of information (the number of bits required to encode it). The particular information that is encoded ought not to matter. But what the mental rotation experiments seem to show is that there are information-processing tasks that take varying amounts of time even though the quantity of information remains the same.

Exercise 2.5 Why does a byte carry 256 units of information? (*Hint:* There are eight digits, each of which can be in one of two states — so how many possible states are there?)

To see the issue, look again at the experimental drawings in Figure 2.7 and think about how each of them might be digitally encoded. Suppose that we think of each drawing as divided into many small boxes (rather like pixels on a television screen or computer monitor). Because the drawings are in black and white, we can convey a lot of information about the drawing by stating, for each pixel, whether it is black or white. But this will not give us a full characterization because the figures are represented three-dimensionally. This means that our characterization of each pixel that represents part of a surface will have to include a value for the surface ′ s degree of orientation, degree of brightness, and so on.

Now, suppose that this has been done and that we have a pixel-by-pixel description of each drawing. Each pixel description is simply a set of numbers that specifies the values on the relevant dimensions at a pixel location.

If this is how information is encoded, then solving the problem is essentially a matter of comparing two numerical descriptions to see if one can be mapped onto the other. Solving this problem is a tricky piece of mathematics that we, fortunately, do not have to go into, but there is no obvious reason why it should take longer for pairs of figures that are at greater degrees of rotation from each other than for pairs that are at smaller degrees from each other — and certainly no reason why there should be a linear relationship between reaction time and degree of rotation.

For reasons such as these, then, many cognitive scientists have suggested that mental rotation tasks tap into ways of encoding information very different from how information is encoded in a digital computer.

One distinctive feature of how information is represented in digital computers (what is often called *digital representation*) is that the connection

between what we might think of as the unit of representation and what that unit represents is completely arbitrary. There is no particular reason to use the symbol " 0 " for a black pixel and " 1 " for a white pixel, rather than the other way around.

Contrast this with how, for example, a map represents a geographical region. There is a large-scale resemblance between the principal geographical features of the region and the discernible features of the map — without it, the map will not be of much use. The weaving and winding of a river are matched by the weaving and winding of the line on the map that represents the river. The outlines of a region of forestry are matched by the edges of the green patch on the map. And so on. A map is an excellent example of an imagistic representation. The basic characteristic of an imagistic representation is that representation is secured through resemblance.

Exercise 2.6 Can you think of other differences between digital representation and imagistic representation?

One popular interpretation of the mental rotation experiments is as showing that at least some types of information are represented imagistically at the level of subconscious information processing. It is not just that we have the experience of consciously rotating figures in our mind ' s eye. The shapes are also represented imagistically in the subconscious information processing that makes possible these types of conscious experience. This is thought to make possible operations that can be carried out on imagistically

represented information but not on digitally represented information — for example, rotation.

The idea that the information processing in mental imagery involves operations on imagistic representations also makes sense of many of the other effects identified in the experimental literature on imagery. For example, in a famous experiment carried out by Stephen Kosslyn in 1973, subjects were asked to memorize a set of drawings like those illustrated in Figure 2.9 .

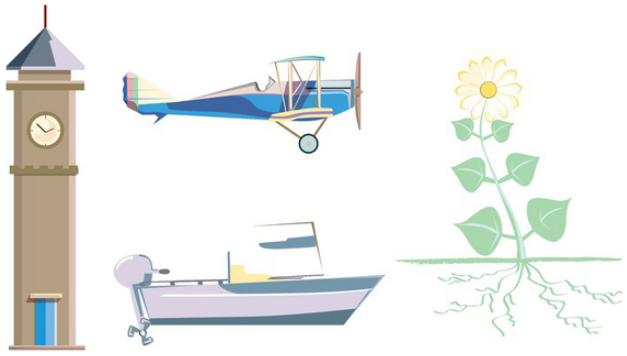


Figure 2.9 Examples of vertically and horizontally oriented objects that subjects were asked to visualize in Kosslyn ' s (1973) scanning study.

(Adapted from Kosslyn et al. 2006 .)

Kosslyn then gave them the name of one of the objects (e.g., " airplane ") and asked them to focus on one end of the memorized drawing. The experiment consisted of giving the subjects the names of possible parts of the object (e.g., " propeller ") and asking them to examine their images to see whether the object drawn did indeed have the relevant part (which it did in 50 percent of the trials). The subjects pushed a button only if they did indeed see the named part in their image of the drawn object.

Kosslyn found that the length of time it took the subjects to answer varied according to the distance of the parts from the point of focus. If the

subjects were asked to focus on the tail of the plane, it would take longer for them to confirm that the plane had a propeller than that there was not a pilot in the cockpit.

Kosslyn ′s interpretation of his own experiment was that the type of information processing used in answering the test questions involves scanning imagistic representations. Instead of searching for the answer within a digitally encoded database of information about the figures, the subjects scan an imagistically encoded mental image of the airplane.

> **Exercise 2.7** Can you think of a way of explaining the results of Kosslyn ′s experiments without the hypothesis of imagistically encoded information?

This takes us to the heart of fundamental issues in cognitive science. Almost all cognitive scientists agree that cognition is information processing. But the imagery debate shows that there are competing models of how information is stored and how it is processed. We will return to these issues in later chapters.

2.3 An Interdisciplinary Model of Vision

The mind can be studied at many different levels. We can study the mind from the bottom up, beginning with individual neurons and populations of neurons, or perhaps even lower down, with molecular pathways whose activities generate action potentials in individual neurons, and then trying to build up from that to higher cognitive functions. Or we can begin from the

top down, starting out with general theories about the nature of thought and the nature of cognition and working downward to investigate how corresponding mechanisms might be instantiated in the brain. Each approach involves distinct levels of explanation that often have separate disciplines corresponding to them. A fundamental problem for cognitive science is working out how to combine and integrate different levels of explanation.

Levels of Explanation: Marr ' s *Vision* (1982)

The earliest systematic approach to tackling this problem is David Marr ' s model of the human visual system in his 1982 book *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* Marr ' s conception of how different levels of explanation connect up with each other has been deeply influential, both as a blueprint for practicing scientists and as a model for understanding the nature of explanation in cognitive science.

Marr distinguishes three different levels for analyzing cognitive systems. At the top is the *computational level* . Here, cognitive scientists analyze, in very general terms, the particular type of task that the system performs. The tasks of an analysis at the computational level are as follows:

1. To translate a general description of the cognitive system into a specific account of the particular information-processing problem that the system is configured to solve

2. To identify the constraints that hold upon any solution to that information-processing task

On the assumption that cognition is ultimately information processing, a computational analysis identifies the information with which the cognitive system has to begin (the *input* to that system) and the information with which it needs to end up (the *output* from that system).

Exercise 2.8 Think of a specific cognitive system and explain what it does in information-processing terms .

Marr ' s next level down is the *algorithmic level* . This level tells us how the cognitive system actually solves the specific information-processing task identified at the computational level. It tells us how the input information is transformed into the output information. It does this by giving algorithms that result in that transformation.

So, an algorithmic-level explanation takes the form of specifying detailed sets of information-processing instructions that will explain how, for example, information from the sensory systems about the distribution of light in the visual field is transformed into a representation of the three-dimensional environment around the perceiver.

In contrast, the principal task at the *implementational level* is to find a physical realization for the algorithm — that is to say, to identify physical structures that will realize the representational states over which the algorithm is defined and to find mechanisms at the neural level that can properly be described as computing the algorithm in question.

Exercise 2.9 Explain in your own words the difference between algorithmic and implementational explanations.

Table 2.1 is a table from Marr ' s book illustrating how the different levels of explanation fit together. This approach is a classic example of *top-down* analysis. It starts with a high-level analysis of the specific information-processing problems that the visual system confronts, as well as the constraints under which it operates. At each stage of the analysis, these problems become more circumscribed and more determinate. Accounts at the algorithmic and implementational levels are motivated by discussions of constraint and function at the computational level — that is, by considering which features of the environment the organism needs to model and the resources it has available to it.

Table 2.1 The three different levels that Marr identified for explaining information-processing systems

COMPUTATIONAL THEORY	REPRESENTATION AND ALGORITHM	HARDWARE IMPLEMENTATION
What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?	*How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?*	*How can the representation and algorithm be realized physically?*

Note. Each level has its own characteristic questions and problems.

(From Marr 1982 .)

Applying Top-Down Analysis to the Visual System

We can get a better sense of how this general model of top-down analysis works in practice by looking at how Marr applied it in thinking about human

vision.

Marr ′ s model is very interdisciplinary. His thinking at the computational level about what the visual system does was strongly influenced by research into patients with brain damage carried out by clinical neuropsychologists. In his book, he explicitly refers to Elizabeth Warrington ′ s work on patients with damage to the left and right parietal cortexes — areas of the brain that tend to produce problems in perceptual recognition when damaged.

Warrington noticed that the perceptual deficits of the two classes of patients are fundamentally different. Patients with right parietal lesions are able to recognize and verbally identify familiar objects *provided that they can see them from familiar or " conventional " perspectives.* From unconventional perspectives, however, these patients would not only fail to identify familiar objects but would also vehemently deny that the shapes they perceived could possibly correspond to the objects that they, in fact, were. Figure 2.10 provides an example of conventional and unconventional perspectives.



Figure 2.10 Two images of a bucket. A familiar/conventional view is on the left, and an unfamiliar/unconventional view is on the right.

(From Warrington and Taylor 1973 .)

Patients with left parietal lesions showed a diametrically opposed pattern. Although left parietal lesions are often accompanied by language problems, patients with such lesions tend to be capable of identifying the shape of objects. They are as successful as normal subjects on matching tasks and are perfectly able to match conventional and unconventional representations of the same object.

Marr drew two conclusions about how the visual system functions from Warrington ′ s neuropsychological observations. First, information about the shape of an object must be processed separately from information about what the object is for and what it is called. Second, the visual system can deliver a specification of the shape of an object even when the perceiver is unable to recognize the object. He explains:

> Elizabeth Warrington had put her finger on what was somehow the quintessential fact about human vision — that it tells us about shape and space and spatial arrangement. Here lay a way to formulate its purpose — building a description of the shapes and positions of things from images. Of course, that is by no means all that vision can do; it also tells us about the illumination and about the reflectances of the surfaces that make the shapes — their brightnesses and colors and visual textures — and about their motion. But these things seemed secondary; they could be hung off a theory in which the main job of vision was to derive a representation of shape.
>
> (Marr 1982 : 7)

So, at the computational level, the visual system ′ s basic task is to construct a representation of the three-dimensional shape and spatial arrangement of an object in a form that will allow that object to be recognized. Because ease

of recognition is correlated with the ability to extrapolate from the particular vantage point from which an object is viewed, Marr concluded that this representation of object shape should be on an object-centered rather than an egocentric frame of reference (where an egocentric frame of reference is one centered on the viewer). This, in essence, is the theory that emerges at the computational level.

Exercise 2.10 Explain in your own words why Marr drew the conclusions he did from Elizabeth Warrington ′ s patients.

Analysis at the algorithmic level calls for a far more detailed account of how the general information-processing task identified at the computational level is carried out. We are looking for an algorithm that can take the system from inputs of the appropriate type to outputs of the appropriate type. This raises a range of new questions. How exactly is the input and output information encoded? What are the system ′ s *representational primitives* (the basic " units " over which computations are defined)? What sort of operations is the system performing on those representational primitives to carry out the information-processing task?

A crucial part of the function of vision is to recover information about surfaces in the field of view — in particular, information about their orientation, how far they are from the perceiver, and how they reflect light. In Marr ′ s theory, this information is derived from a series of increasingly complex and sophisticated representations, which he terms the *primal sketch* , the *2.5-dimensional (2.5D) sketch* , and the *three-dimensional (3D) sketch.*

The primal sketch makes explicit some basic types of information implicitly present in the retinal image. These include distributions of light intensity across the retinal image — areas of relative brightness or darkness, for example. It also represents the basic geometry of the field of view. Figure 2.11 gives two illustrations. Note how the primal sketch reveals basic geometrical structure — an embedded triangle in the left figure and an embedded square in the right.

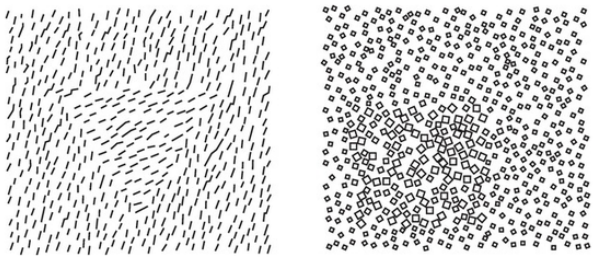


Figure 2.11 Two examples of Marr ′ s primal sketch, the first computational stage in his analysis of the early visual system. The primal sketch contains basic elements of large-scale organization (the embedded triangle in the left-hand sketch, for example).

(Adapted from Marr 1982 .)

The next task is to extract from the primal sketch information about the depth and orientation of visible surfaces from the viewer ′ s perspective. This yields the 2.5D sketch. For every point in the field of view, the 2.5D sketch represents how far it is from the observer, as illustrated in Figure 2.12 .

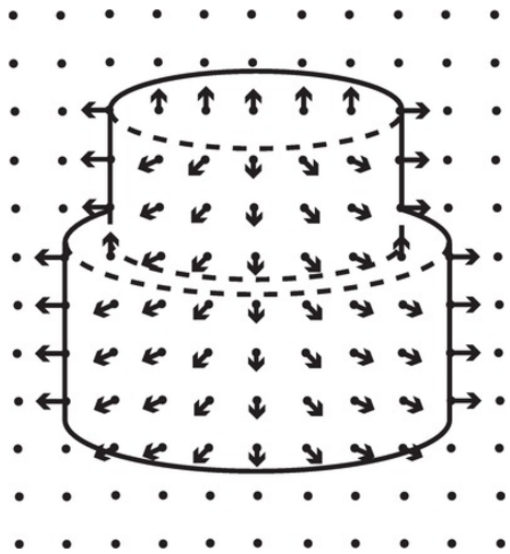


Figure 2.12 An example of part of the 2.5D sketch. The figure shows orientation information but no depth information.

(Adapted from Marr 1982 .)

The 2.5D sketch is viewer centered. It depends on the viewer ′ s particular vantage point. However, the visual system allows us to keep track of objects even though their visual appearance changes from the viewer ′ s perspective (because either the object or the viewer is moving, for example). This requires a stable representation of object shape that is independent of the viewer ′ s particular viewpoint. This viewer-independent representation is provided by the 3D sketch, as illustrated in Figure 2.13 .

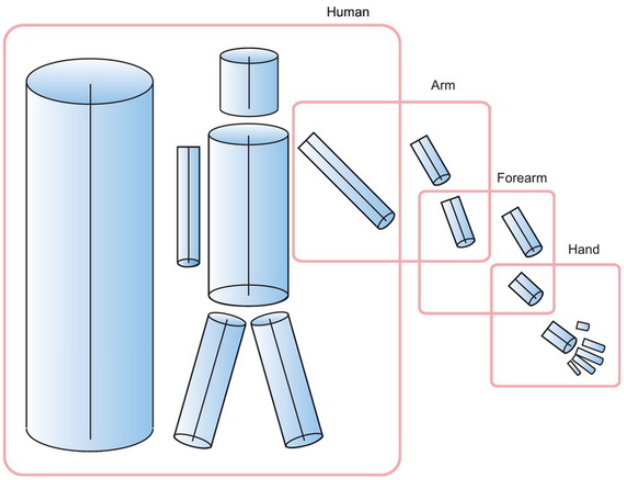


Figure 2.13 An illustration of Marr ′ s 3D sketch, showing how the individual components are constructed. The 3D sketch gives an observer-independent representation of object shape and size.

(Adapted from Marr 1982 .)

At the algorithmic level, the job is to specify these different sketches and explain how the visual system gets from one to the next, starting with the basic information arriving at the retina. Because the retina is composed of cells that are sensitive to light, this basic information is information about the intensity of the light reaching each of those cells .

What are the starting points for the information processing that will yield as its output an accurate representation of the layout of surfaces in the distal environment? Marr ′ s answer is that the visual system needs to start with discontinuities in light intensity because these are a good guide to boundaries between objects and other physically relevant properties. Accordingly, the representational primitives that he identifies are all closely correlated with changes in light intensity.

These representational primitives include *zero-crossings* (registers of sudden changes in light intensity), blobs, edges, segments, and boundaries.

The algorithmic description of the visual system takes a representation formulated in terms of these representational primitives as the input and spells out a series of computational steps that will transform this input into the desired output, which is a representation of the 3D perceived environment.

Moving down to the implementational level, a further set of disciplines comes into play. In thinking about the cognitive architecture within which the various algorithms computed by the visual system are embedded, we will obviously need to take into account the basic physiology of the visual system — and this, in turn, is something that we will need to think about at various different levels. Marr ' s own work on vision contains relatively little discussion of neural implementation. But the table from his book shown here as Figure 2.14 illustrates where the implementational level fits into the overall picture. Figure 2.15 is a more recent attempt at identifying the neural structures underlying the visual system.

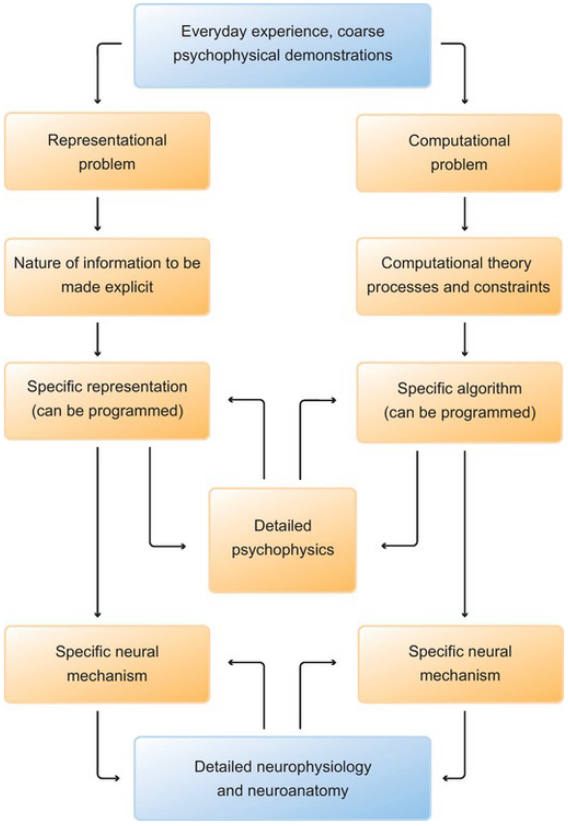


Figure 2.14 The place of the implementational level within Marr ' s overall theory. Note also the role he identifies for detailed experiments in psychophysics (the branch of psychology studying how perceptual systems react to different physical stimuli).

(Adapted from Marr 1982 .)

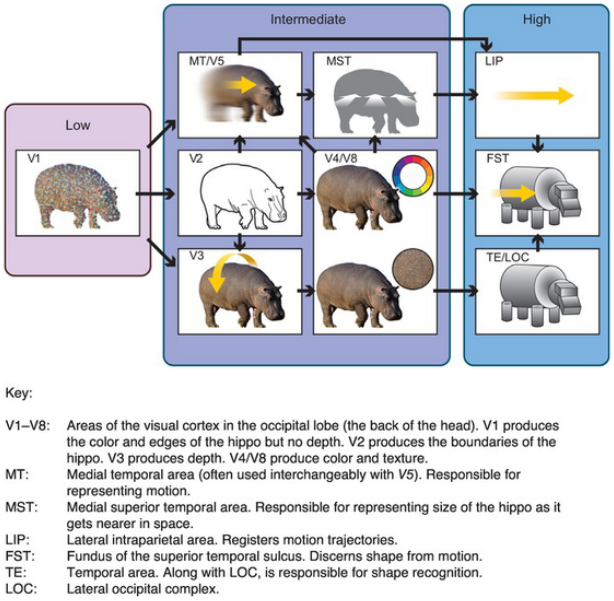


Figure 2.15 An illustration of the hierarchical organization of the visual system, including which parts of the brain are likely responsible for processing different types of visual information.

(From Prinz 2012 .)

Marr ' s analysis of the visual system clearly illustrates both how a single cognitive phenomenon can be studied at different levels of explanation and how the different levels of explanation can come together to provide a unified analysis. It is not surprising that Marr ' s analysis of the visual system has been taken as a paradigm of how cognitive science ought to proceed .

Summary

This chapter continued our historical overview. We reviewed three milestones in the emergence of cognitive science: Terry Winograd ' s SHRDLU program for modeling natural language understanding, the

explorations into the representational format of mental imagery inspired by the mental rotation experiments of Roger Shepard and others, and the multilevel analysis of the early visual system proposed by David Marr. Each illustrates how researchers brought together some of the basic tools discussed in Chapter 1 and applied them to specific cognitive capacities.

Checklists

Winograd ' s SHRDLU

1. SHRDLU is more sophisticated than a conversation-simulating chatbot because it uses language to report on the environment and to plan action.

2. SHRDLU illustrates how abstract grammatical rules can be represented in a cognitive system and integrated with other types of information about the environment.

3. The design of SHRDLU illustrates a common strategy in cognitive science, namely, analyzing a complex system by breaking it down into distinct components, each performing a circumscribed information-processing task.

4. These information-processing tasks are implemented algorithmically (as illustrated by the flowcharts that Winograd used to explain SHRDLU ' s different procedures).

The Imagery Debate

1. The experiments that gave rise to the imagery debate forced cognitive scientists to become much more reflective about information and information processing.

2. The imagery debate is not a debate about conscious experiences of mental imagery. It is about the information processing underlying those conscious experiences.

3. The mental rotation and scanning experiments were taken by many cognitive scientists to show that some information processing involves operations on geometrically encoded representations.

4. The debate is about whether the different effects revealed by experiments on mental imagery can or cannot be explained in terms of digital information-processing models.

Marr ' s Theory of Vision

1. Marr identified three different levels for analyzing cognitive systems.

2. His analysis of vision is a classic example of the top-down analysis of a cognitive system. The analysis is driven by a general characterization at the computational level of the information-processing task that the system is carrying out.

3. This general analysis at the computational level is worked out in detail at the algorithmic level, where Marr explains how the information-processing task can be algorithmically carried out.

4. The bottom level of analysis explains how the algorithm is actually implemented. It is only at the implementational level that neurobiological considerations come directly into the picture.

Further Reading

The general historical works mentioned at the end of the previous chapter also cover the material in this chapter and will provide further useful context-setting.

A web-based version of ELIZA can be found in the online resources. The principal resource for SHRDLU is Winograd ' s book *Understanding Natural Language* (1972). This is very detailed, however, and a more accessible treatment can be found in his article " A Procedural Model of Language Understanding " (Winograd 1973), which is reprinted in Cummins and Cummins 2000 . SHRDLU was an early example of an expert system. A helpful introduction is the entry on expert systems in the *Macmillan Encyclopedia of Cognitive Science* (Medsker and Schulte 2003). The online *Encyclopedia of Cognitive Science* (Nadel 2005) also has an entry on SHRDLU.

Many important articles in the imagery debate are collected in Block 1981 . The experiments described in the text were originally reported in Shepard and Metzler 1971 , Kosslyn 1973 , and Cooper and Shepard 1973 . Demonstrations and further discussion of mental imagery can be found in the online resources. Meta-analyses of mental imagery studies can be found in Voyer et al. 1995 and Zacks 2008 . Kosslyn et al. 2006 and Pearson and Kosslyn 2015 are more recent contributions from central figures in the debate. Zhao and Sala 2016 explore representation strategies in mental

rotation tasks. On mental scanning in general, Saul Sternberg has suggested that subjects search through lists of, say, numbers using a serial scanning process (see Sternberg 2016 for a recent review).

The imagery debate has received considerable attention from philosophers and cognitive scientists interested in the format of mental representations. Rollins 1989 and Tye 1991 are book-length studies. The online *Stanford Encyclopedia of Philosophy* also has an entry on mental imagery. Zenon Pylyshyn has been a prominent supporter of the null hypothesis that mental imagery involves exactly the same representational format as other types of reasoning. See Pylyshyn 2002 and the peer commentary there. For more general discussions of representational formats in perception and imagery, see Quilty-Dunn 2020 and Palmiero, Piccardi, et al. 2019 for very different approaches from the embodied-cognition framework.

Marr ' s book on vision (1982) has recently been reprinted (2010). Shimon Ullman ' s foreword in the new edition and Tomaso Poggio ' s afterword provide useful background and discussion. The first chapter of Marr ' s book is reprinted in Bermúdez 2006 and Cummins and Cummins 2000 , and Marr ' s selected papers are reprinted in Vaina 1991 . Dawson 1998 is a textbook on cognitive science that is structured entirely around Marr ' s tri-level hypothesis. Also see Tsotsos 2011 . Chapter 2 of Prinz 2012 gives a general assessment of the accuracy of Marr ' s account, in light of current research on visual processing. Elizabeth Warrington ' s classic studies can be found in Warrington and Taylor 1973 , 1978 . The journal *Topics in Cognitive Science* published a special issue on Marr in 2015 (available open access). Ballard and Zhang 2021 describes how Marr ' s tri-

level framework has been adapted in the fields of machine vision and robotics.

Chapter 3

The Turn to the Brain

◈

Overview

3.1 Cognitive Systems as Functional Systems?

3.2 The Anatomy of the Brain and the Primary Visual Pathway

The Two-Visual-Systems Hypothesis: Ungerleider and Mishkin, "Two Cortical Visual Systems" (1982)

3.3 Extending Computational Modeling to the Brain

A New Set of Algorithms: Rumelhart et al., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (1986)

Pattern Recognition in Neural Networks: Gorman and Sejnowski, "Analysis of Hidden Units in a Layered Network Trained to Identify Sonar Targets" (1998)

3.4 Mapping the Stages of Lexical Processing

Functional Neuroimaging with PET

Petersen et al., "Positron Emission Tomographic Studies of the Cortical Anatomy of Single-Word Processing" (1988)

3.5 Studying Memory for Visual Events

Functional Neuroimaging with fMRI

Brewer et al., " Making Memories: Brain Activity That Predicts How Well Visual Experience Will Be Remembered " (1998)

3.6 **The Neural Correlates of the BOLD Signal**

Logothetis, " The Underpinnings of the BOLD Functional Magnetic Resonance Imaging Signal " (2001)

Overview

A striking feature of contemporary cognitive science, as compared with the 1970s, for example, is the increasing centrality of the brain. This chapter reviews some landmarks in cognitive science ' s " turn to the brain. "

For both theoretical and practical reasons, neuroscience was fairly peripheral to cognitive sciences until the 1980s. Section 3.1 looks at some of the theoretical reasons, particularly the influential idea that cognitive systems are functional systems and thus need to be studied in terms of their function — what they do and how they do it. Many cognitive scientists hold that this type of functional analysis can be carried out without looking at the details of neural implementation.

One early move away from this functional view came with the *two-visual-systems hypothesis* , originally proposed by neuroscientists Leslie Ungerleider and Mortimer Mishkin. Section 3.2 relates how Ungerleider and Mishkin drew conclusions about the structure and organization of vision from data about the pathways in the brain that carry visual information. The direction of explanation is from the bottom up rather than the top down (as in Marr ' s framework, which we looked at in Section 2.3).

An important factor in the turn toward the brain was the development of ways of modeling cognitive abilities designed to reflect very general properties of brains. As described in Section 3.3 , so-called *connectionist networks* , or *artificial neural networks* , involve large populations of neuronlike units. The individual units are not biologically plausible in any detailed sense. But the network as a whole behaves in ways that reflect certain high-level properties of brain functioning.

Until the 1980s, techniques for studying human brains while cognitive tasks were actually being carried out were relatively unsophisticated and not widely known among cognitive scientists. This changed with the emergence of functional neuroimaging in the 1980s, which provided a powerful tool for studying what goes on in the brain when subjects are actually performing different types of cognitive tasks.

In Section 3.4 , we look at an early and very influential positron emission tomography (PET) study of visual word processing. This study shows how functional neuroimaging can be used to generate information-processing models of how cognitive tasks are carried out — information-processing models that are derived not from abstract task analysis but, rather, from the detailed study of neural activity.

Section 3.5 introduces functional magnetic resonance imaging (fMRI), which has superseded PET in many domains and allows a different type of experimental design (known as *event-related design*). Event-related fMRI is introduced through studies of visual memory.

Finally, in Section 3.6 , we turn to what fMRI actually measures, looking at groundbreaking experiments by Nikos Logothetis. These experiments use single-electrode recordings of individual neurons to study

the type of brain activity that is correlated with the blood-oxygen-level – dependent (BOLD) signal that is directly measured by fMRI.

3.1 Cognitive Systems as Functional Systems?

First, though, some background. Most of the models that we have looked at so far in our historical survey assume that information is transformed and transmitted in the brain in much the same way as information is transformed and transmitted in digital computers. Because we can study computer algorithms without thinking about the hardware and circuitry on which they run, it is not surprising that these models typically abstract away from the details of neural machinery.

In fact, many cognitive scientists think cognitive processes *have to* be studied independently of the neural machinery on which they run. They think of cognitive systems as *functional* systems and, hence, that they should be understood primarily in terms of their function — what they do and how they do it. And, these cognitive scientists emphasize, this type of analysis is independent of the particular physical structure implementing that function.

An analogy will help. Consider a heart. Hearts are organs that pump blood around the body. They collect deoxygenated blood and pump it toward the lungs, where it becomes reoxygenated. The actual physical structure of the heart is not particularly important. An artificial heart will do the job just as well (although not perhaps for as long) and so still counts as a heart. Crocodiles and humans have hearts with four chambers, whereas

most reptiles have hearts with three chambers. What matters is the job the heart does, not how it does it. A gray whale ′s heart is no less a heart than a hummingbird ′s heart just because the first beats 9 times per minute while the second beats 1,200 times. One way of putting this is to say that functional systems are *multiply realizable*. The heart function can be realized by multiple different physical structures.

Exercise 3.1 Give another example of a multiply realizable system.

If cognitive systems are multiply realizable functional systems like the heart, then it is a mistake to concentrate on biological details of how the brain works. In fact, some cognitive scientists think, this may lead to a misleading picture of how cognition works. It might lead us to consider as essential to memory, say, things that are really just contingent properties of how our brains have evolved. We would be making the same mistake as if we were to conclude that hearts need to have four chambers because the human heart does, or if we decided that Microsoft Word has to run on an Apple M1 chip just because that is the processor in my Mac.

However, other cognitive scientists think that abstracting away from neural machinery in studying the algorithms of cognition may be misleading. For one thing, cognitive activity needs to be coordinated with behavior and adjusted online in response to perceptual input, both of which require cognitive systems with a highly developed sense of timing. The right answer is no use if it comes at the wrong time. But how can we think about the speed and efficiency of the mind without taking into account the fact that it runs on the hardware of the brain?

Moreover, cognitive abilities and skills themselves emerge over time. Eventually, they deteriorate and, for many of us, gradually fade out of existence. In some unfortunate cases, they are drastically altered as a result of traumatic damage. This means that an account of the mind must be compatible with plausible accounts of how cognitive abilities emerge. It must be compatible with what we know about how cognitive abilities deteriorate. It must be compatible with what we know about the relation between damage to the brain and cognitive impairment.

Cognitive abilities tend to *degrade gracefully*. As we get older, reaction times increase, motor responses slow down, and recall starts to become more problematic. But these abilities do not (except as a result of trauma or disease) disappear suddenly. The deterioration is gradual, incremental, and usually imperceptible within small time frames. This type of graceful degradation is a function of how brains are wired and of the biochemistry of individual neurons.

The same holds for how cognitive abilities emerge and develop. Brains learn the way they do because of how they are constructed — and in particular because of the patterns of connectivity existing at each level of neural organization (between neurons, populations of neurons, neural systems, neural columns, and so forth). These facts place constraints on our higher-level theories of cognitive abilities.

Exercise 3.2 Can you think of other reasons for thinking that we should not theorize about cognition without theorizing about the brain?

3.2 The Anatomy of the Brain and the Primary Visual Pathway

In order to appreciate the two-visual-systems hypothesis, as our first illustration of the turn to the brain, we need a little information about the large-scale anatomy of the brain.

Anatomists distinguish three different parts of the mammalian brain — the *forebrain* , the *midbrain* , and the *hindbrain.* This structure is illustrated for the human brain in Figure 3.1 .

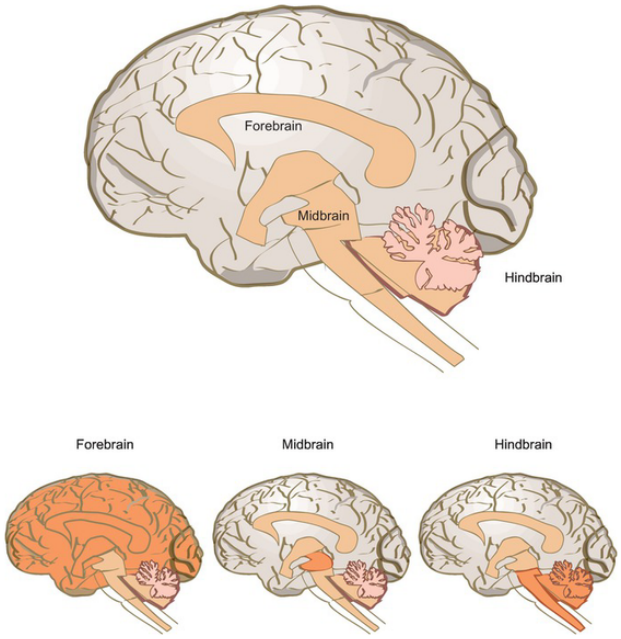


Figure 3.1 The large-scale anatomy of the brain, showing the forebrain, the midbrain, and the hindbrain.

As the figure shows, the forebrain is the largest region. Most of it is taken up by the *cerebrum* (see Figure 3.2), which is important for cognitive and motor processing. The cerebrum is divided into left and right

hemispheres, separated by a deep groove (the *longitudinal* or *interhemispheric fissure*).

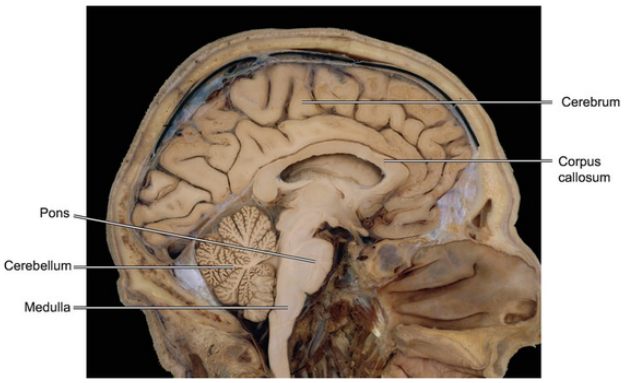


Figure 3.2 A vertical slice of the human brain, showing the cerebrum.

(© TISSUEPIX/SCIENCE PHOTO LIBRARY .)

The two hemispheres have similar organizations. Each has an outer layer, the cerebral cortex. Only mammals have a cerebral cortex, and in the human brain it is about 2 – 4 mm thick. Moving inward, we find the remaining major structures of the forebrain. These are the thalamus, the hypothalamus, and the limbic system, collectively known as *subcortical areas* (because they lie below the cortex).

The tissue of these inner, subcortical parts of the forebrain is known as *white matter* because it is made up of neurons whose axons are encased in a white sheath of myelin (which speeds up the transmission of nerve signals). Neurons in the cerebral cortex are typically unmyelinated and look gray — hence the term *gray matter* .

Within each hemisphere, the cerebral cortex is divided into four main regions, called *lobes.* Each lobe is believed to be responsible for carrying out different cognitive tasks. Figure 3.3 illustrates the organization of the

left hemisphere into four lobes, and Box 3.1 summarizes what each lobe is believed to be specialized for.

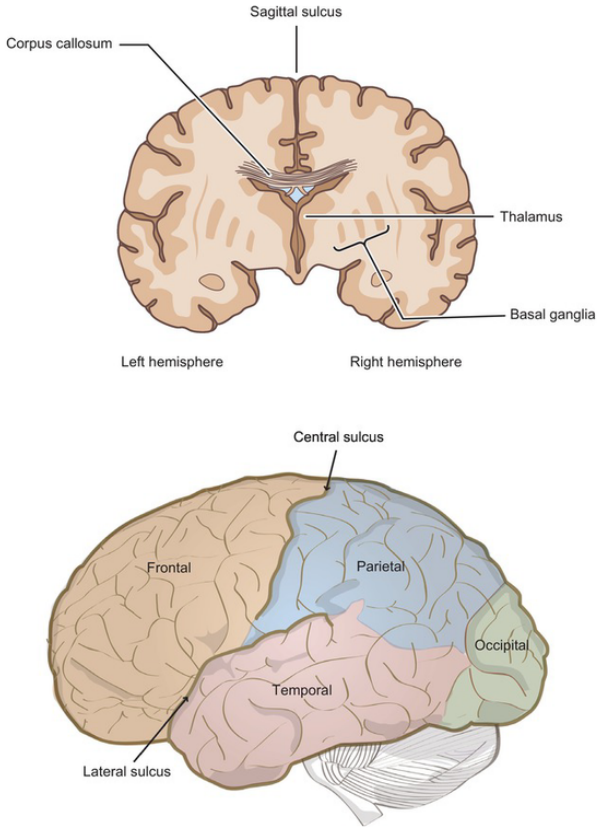


Figure 3.3 The division of the left cerebral hemisphere into lobes.

(From Stein and Stoodley 2006 .)

BOX 3.1 What Does Each Lobe Do?

- Frontal lobe — reasoning, planning, parts of speech, movement, emotions, and problem solving
- Parietal lobe — movement, orientation, recognition, perception of stimuli
- Occipital lobe — associated with visual processing
- Temporal lobe — associated with perception and recognition of auditory stimuli, memory, and speech

The lobes are themselves highly organized. In 1909, German neurologist Korbinian Brodmann proposed a mapping of the cerebral cortex into fifty-two areas. These Brodmann areas are still in use today. An example particularly relevant to us now is Brodmann area 17, which is also known as the *primary visual cortex*, the *striate cortex*, or *area V1*. Brodmann area 17 is located in the *occipital lobe*, and as the name *primary visual cortex* suggests, it is the point of arrival in the cortex for information from the retina.

The information pathway leading from the retina to the primary visual cortex is well understood and illustrated in Figure 3.4. Visual information from each eye is transmitted by the optic nerve to the lateral geniculate nucleus (a subcortical area of the forebrain) and thence to the primary visual cortex. The diagram clearly shows the *contralateral* organization of the brain. Each hemisphere processes information deriving from the opposite side of space. So, visual information from the right half of the

visual field is processed by the left hemisphere (irrespective of which eye it comes from).

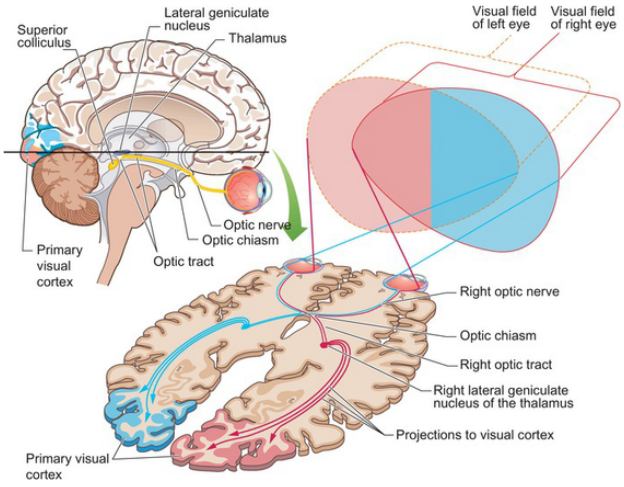


Figure 3.4 The primary visual pathway. Note the contralateral organization, with information from the right side of space processed by the left side of the brain.

(From Stein and Stoodley 2006 .)

Much less well understood is what happens to that information when it leaves the primary visual cortex. This is where we come to the two-visual-systems hypothesis and the work of Ungerleider and Mishkin.

The Two-Visual-Systems Hypothesis: Ungerleider and Mishkin, " Two Cortical Visual Systems " (1982)

The two-visual-systems hypothesis is important both because of the tools that were used to arrive at it (including the study of patients with brain damage and experiments on monkeys) and because it illustrates a bottom-up, as opposed to top-down, way of studying the mind.

Ungerleider and Mishkin suggested that the specific route information takes from the primary visual cortex depends on the type of information it is. Information relevant to recognizing and identifying objects follows a *ventral route* (see Box 3.2) from the primary visual cortex to the temporal lobe, whereas information relevant to locating objects in space follows a *dorsal* route from the primary visual cortex to the posterior parietal lobe. The two routes are illustrated in Figure 3.5 .

BOX 3.2 Brain Vocabulary

Neuroscientists and neuroanatomists use an unfamiliar vocabulary for talking about the layout of the brain:

Rostral = at the front

Caudal = at the back

Ventral = at the bottom

Dorsal = at the top

Ipsilateral = same side

Contralateral = opposite side

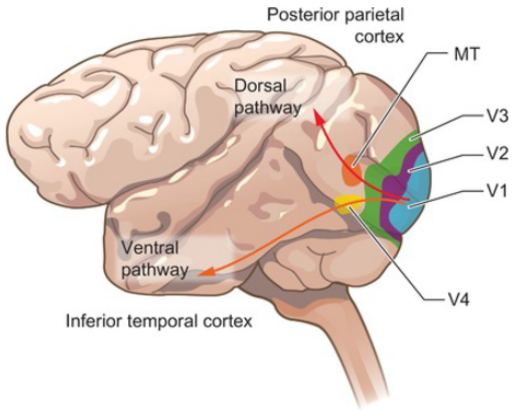


Figure 3.5 Image showing ventral (*orange*) and dorsal (*red*) pathways in the human visual system.

Ungerleider and Mishkin came to this conclusion after studying cognitive impairments due to brain damage and performing neuroanatomical experiments on monkeys. By the time they were writing, there was already considerable evidence from patients with brain damage that damage to the temporal and parietal lobes produced very different types of cognitive problems. Damage to the temporal cortex is associated with problems in identifying and recognizing objects, whereas damage to the parietal cortex tends to result in problems locating objects.

Evidence of this type has always been very important in working out the function of the different lobes (see Box 3.1 for a standard " division of labor " between the lobes). But it does not explain the actual path that information takes in the brain.

Simply finding specific cognitive problems associated with damage to a specific brain region gives us no way of telling whether the impaired cognitive abilities are normally carried out by the damaged brain region itself or by some other brain region that crucially depends on input from the damaged brain region. Solving this problem cannot be done simply by

observing the results of brain damage. Precise surgical intervention is required, in the form of targeted removal of specific brain areas to uncover the connections between them.

Ungerleider and Mishkin turned to *cross-lesion disconnection experiments* on monkeys. These experiments exploit the fact that the cerebrum is divided into two hemispheres, with duplication of the principal cortical areas. Ungerleider and Mishkin hypothesized that there is a pathway connecting the primary visual cortex and the inferior temporal area, and so they performed surgery in monkeys to remove the primary visual cortex from one hemisphere in monkeys and the inferior temporal area from the other hemisphere. This destroyed the postulated pathway in each hemisphere. However, because the hemispheres can communicate through a large bundle of fibers known as the *corpus callosum* (illustrated in Figure 3.2), it turned out that there was little or no loss of function in the monkeys — until the corpus callosum was severed to stop information flowing between the hemispheres.

Using experiments such as these (in addition to other types of neurophysiological evidence), Ungerleider and Mishkin conjectured that information relevant to object identification and recognition flows from the primary visual cortex to the inferior temporal cortex via areas in the occipital lobe collectively known as the *prestriate cortex.* They called this the *ventral pathway* .

They identified a completely different pathway (the *dorsal pathway*) leading from the primary visual cortex to the posterior parietal lobe. Once again, they used cross-lesion disconnection experiments. In this case the task was the so-called *landmark task* , illustrated in the top-left part of Figure 3.6 . In the landmark task, monkeys are trained to choose food from

one of two covered food wells, depending on which is closer to a striped cylinder, which is moved at random.

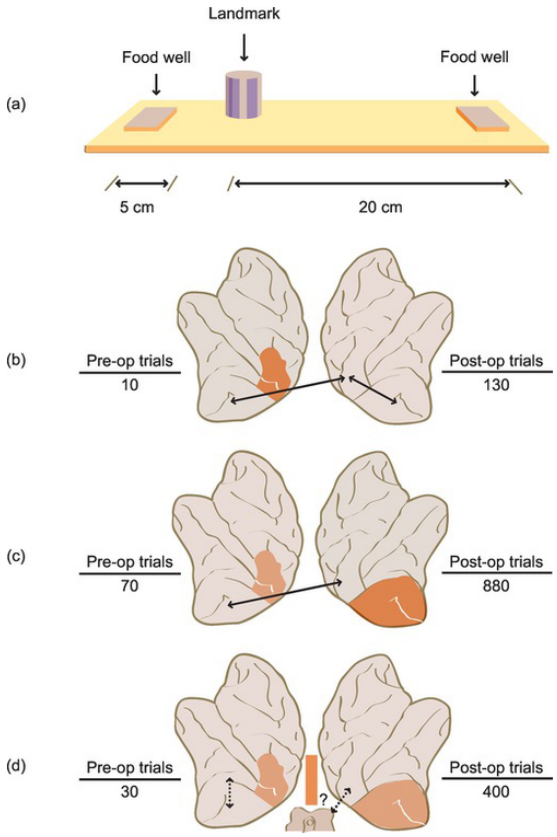


Figure 3.6 Design and results of Ungerleider and Mishkin ′s cross-lesion disconnection studies. (a) Landmark task. Monkeys were rewarded for choosing the covered food well located closer to a striped cylinder (the " landmark "), which was positioned on the left or the right randomly from trial to trial but always 5 cm from one food well and 20 cm from the other. Training was given for 30 trials per day to a criterion of 90 correct responses in 100 consecutive trials. (b) Discrimination retention before and after first-stage lesion (unilateral posterior parietal); 10 preoperative trials and 130 postoperative trials. (c) Discrimination retention before and after second-stage lesion (contralateral striate); 70 preoperative and 880 postoperative trials. (d) Discrimination retention before and after third-stage lesion (corpus callosum); 30 preoperative and 400 postoperative trials. At each stage, the lesion is shown in dark brown

and the lesions of prior stages in light brown. Arrows denote hypothetical connections left intact by lesions.

(Adapted from Cummins and Cummins 2000 .)

The disconnection surgery proceeded in three stages. In the first stage (b in Figure 3.6), the posterior parietal cortex was removed from one side. The second stage (c) removed the primary visual cortex on the opposite side. The final stage (d) was a *transection* (severing) of the corpus callosum.

As indicated in Figure 3.6 , the monkeys were tested on the landmark task both before and after each surgery. The numbers in Figure 3.6 indicate the number of trials required to train the monkeys to a 90 percent success rate on the landmark task. So, for example, prior to the first stage of the surgery, the average number of training trials required was ten. After a lesion of the posterior parietal cortex, the number of training trials went up to 130.

Interestingly, the most severe impairment was caused by the second stage in the surgery, the removal of the primary visual cortex (in contrast to the other experiments on the visual recognition pathway, where severe impairments appeared only with the cutting of the corpus callosum). Ungerleider and Mishkin concluded from this that the posterior parietal cortex in a given hemisphere does not depend much on information about the ipsilateral visual field (see Box 3.2) from the opposite hemisphere ' s primary visual cortex.

This raises an intriguing possibility because it is known that each hemisphere is specialized for the contralateral region of space. It may be that the posterior parietal cortex in each hemisphere is specialized for

processing information about the opposite region of space — so that, for example, the left posterior parietal cortex processes information about the layout of space on the perceiver ' s right-hand side.

This could shed light on the neurological disorder of *unilateral spatial neglect.* Patients with this disorder typically " neglect " one-half of the space around them, eating food from only one side of the plate and describing themselves as unaware of stimuli in the neglected half of space. Unilateral spatial neglect typically follows damage to the posterior parietal cortex in one hemisphere (most often the right), and the neglected region is contralateral to the damage (so that, most often, the left-hand side of space is neglected).

The two-visual-systems hypothesis was a very important step in mapping out the *connectivity* of the brain. Ungerleider and Mishkin ' s basic distinction between the " what " system (served by the ventral pathway) and the " where " system (served by the dorsal pathway) has been refined and modified by many researchers (see the references in Further Reading). However, the idea that there is no single pathway specialized for the processing of visual information and that visual information instead takes different processing routes depending on what type of information it is has proved very enduring.

From the perspective of cognitive science, the significance of the two-visual-systems hypothesis is that it exemplifies, in a particularly clear way, the bottom-up study of how information is processed in the mind. Instead of abstracting away from the details of the channels and pathways between neural systems along which information processing flows, Ungerleider and Mishkin started with those channels and pathways and worked upward

toward identifying distinct cognitive systems carrying out distinct cognitive functions .

Exercise 3.3 Make as detailed a list as you can of similarities and differences between the top-down and bottom-up approaches to studying the organization of the mind.

3.3 Extending Computational Modeling to the Brain

The historical antecedents of neurally inspired computational models go back to the 1940s, particularly to the work of Warren McCullogh and Walter Pitts (discussed in more detail in Chapter 5). But an important modern landmark was the publication in 1986 of a very influential two-volume collection of papers by Rumelhart, McClelland, and the PDP Research Group.

A New Set of Algorithms: Rumelhart et al., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (1986)

The papers in this collection proposed and pursued a new set of abstract mathematical tools for modeling cognitive processes. These models are sometimes called *connectionist networks* and sometimes *artificial neural networks* (terms that we will use interchangeably). They abstract away from many biological details of neural functioning. But they manage to capture some of the key features of how the brain works. We will be looking at

artificial neural networks in much more detail in later chapters (particularly Chapter 5). Here, we will simply give a brief sketch of some key features.

First, connectionist networks display *parallel processing*. An artificial neural network contains a large number of units (artificial neurons). Each unit has a varying level of activation, typically represented by a real number between − 1 and 1. The units are organized into layers, with the activation value of a given layer determined by the activation values of all the individual units. The simultaneous activation of these units, and the consequent spread of activation through the layers of the network, governs how information is processed within the network. The processing is parallel because the flow of information through the network is determined by what happens in all of the units in a given layer — but those units are not connected to each other.

Second, each unit in a given layer has connections running to it from units in the previous layer (unless it is a unit in the input layer) and will have connections running forward to units in the next layer (unless it is a unit in the output layer). The pattern of connections running to and from a given unit is what identifies that unit within the network. The strength of the connection (the *weight* of the connection) between individual neurons varies. In fact, neural networks learn by modifying their weights.

Third, there are no intrinsic differences between one unit and another. The differences lie in the connections holding between that unit and other units.

Finally, most artificial neural networks are trained rather than programmed. They are generally constructed with broad, general-purpose learning algorithms that work by changing the connection weights between units in a way that eventually yields the desired outputs for the appropriate

inputs. These algorithms work by changing the weights of the connections between pairs of neurons in adjacent layers in order to reduce the " mistakes " that the network makes. Figure 3.7 illustrates a generic neural network with three layers of units. The first layer is made up of input units, which receive inputs from sources outside the network. The third layer is made up of output units, which send signals outside the network. The middle layer is composed of what are called *hidden units* .

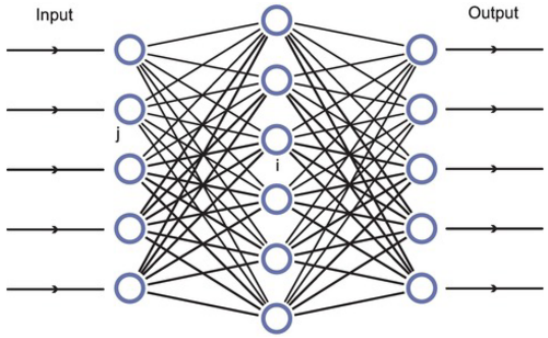


Figure 3.7 A generic three-layer connectionist network (also known as an *artificial neural network*). The network has one layer of hidden units.

(Adapted from McLeod et al. 1998 .)

Hidden units only communicate with units within the network. They are the key to the computational power of artificial neural networks. Networks without hidden units can only perform very limited computations. The network illustrated in Figure 3.7 only has one layer of hidden units, but in fact, networks can be constructed with as many layers as required. (More details coming up in Chapter 5 .)

It takes a long time to train a network. Typically, the network starts with randomly assigned weights. It is then given a training series of input patterns of activation, each of which is associated with a target output pattern of activation. The input patterns are presented. Differences between

the actual output pattern and the target output pattern result in changes to the weights. (This is what the learning algorithm does — it adjusts the weights in order to reduce the difference between actual and desired output.)

This training process continues until errors have diminished almost to zero, resulting in a distinctive and stable pattern of weights across the network. The overall success of a network can be calculated by its ability to produce the correct response to inputs on which it has not been trained. The next subsection illustrates the sort of task that a network can be trained to do with a justly celebrated example .

Pattern Recognition in Neural Networks: Gorman and Sejnowski, " Analysis of Hidden Units in a Layered Network Trained to Identify Sonar Targets " (1998)

Artificial neural networks are particularly suited for pattern-recognition tasks. Here is a classic example. Consider the task of identifying whether a particular underwater sonar echo comes from a submerged mine or from a rock. There are discriminable differences between the sonar echoes of mines and rocks, but there are equally discriminable differences between the sonar echoes from different parts of a single mine or from different parts of a single rock. It is no easy matter to identify reliably whether a sonar echo comes from a mine or from a rock. Human sonar operators can do so reasonably well (after a considerable amount of practice and training), but it turns out that artificial neural networks can perform significantly better than humans.

The first problem is coding the external stimulus as a pattern of activation values. The external stimuli are sonar echoes from similarly

shaped and sized objects known to be either mines or rocks. In order to " transform " these sonar echoes into a representational format suitable for processing by the network, the sonar echoes are run through a spectral analyzer that registers their energy levels at a range of different frequencies. This process gives each sonar echo a unique " fingerprint " to serve as input to the network. Each input unit is dedicated to a different frequency, and its activation level for a given sonar echo is a function of the level of energy in the relevant sonar echo at that frequency. This allows the vector of activation values defined over the input units to reflect the unique fingerprint of each sonar echo.

The neural network developed by Paul Gorman and Terrence Sejnowski to solve this problem contains sixty input units, corresponding to the sixty different frequencies at which energy sampling was carried out, and one layer of hidden units. Because the job of the unit is to classify inputs into two groups, the network contains two output units — in effect, a rock unit and a mine unit. The aim of the network is to deliver an output activation vector of <1,0> in response to the energy profile of a rock and <0,1> in response to the energy profile of a mine. Figure 3.8 is a diagrammatic representation of Gorman and Sejnowski ' s mine/rock network.

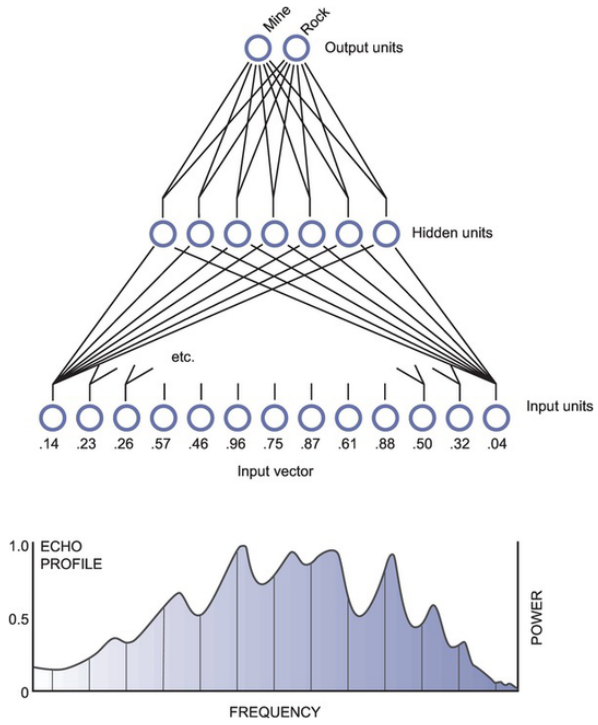


Figure 3.8 Gorman and Sejnowski ' s mine-/rock-detector network.

(Adapted from Gorman and Sejnowski 1988 .)

The mine-detector network is a standard feedforward network (which means that activation is only ever spread forward through the network) and is trained with the *backpropagation learning algorithm* (explained in Chapter 5). Although the network receives information during the training phase about the accuracy of its outputs, the only memory it has of what happened in early sessions is the particular patterns of weights holding across the network. Each time the network comes up with a wrong output (a pattern of <0.83, 0.17> rather than <1.0>, for example, in response to a rock profile), the error is propagated backward through the network, and the weights are adjusted to reduce the error. Eventually, the error at the output units diminishes to a point where the network can generalize to new activation patterns with a 90 percent level of accuracy.

The mine-/rock-detection task is a paradigm of the sort of task for which neural networks are best known and most frequently designed. The essence of a neural network is pattern recognition. But many different types of cognitive ability count as forms of pattern recognition, and the tools provided by artificial neural networks have been used to model a range of cognitive processes — as well as many phenomena that are not cognitive at all (such as predicting patterns in the movements of prices on the stock markets, valuing bonds, and forecasting demand for commodities).

> **Exercise 3.4** Give examples of cognitive abilities that you think would lend themselves to being modeled by artificial neural networks.

3.4 Mapping the Stages of Lexical Processing

We turn now to a different set of techniques that has become an increasingly important part of the cognitive scientist ′ s toolkit.

Functional Neuroimaging with PET

Functional neuroimaging allows brain activity to be studied noninvasively. No surgery is required, and subjects can be studied while they are actually performing experimental tasks.

There are different types of functional neuroimaging. The first experiments we will look at use the technique known as *positron emission*

tomography (better known by its acronym, PET). We will be looking at fMRI (*functional magnetic resonance imaging*) in the next section.

The basic idea behind PET technology (as with functional neuroimaging in general) is to study the function of different brain areas by measuring blood flow in the brain. We can work out which brain areas are involved in carrying out particular cognitive tasks by identifying the areas to which blood is flowing. The distinctiveness of PET is that it provides a safe and precise way of measuring short-term blood flow in the brain. Subjects are given (typically by injection) a small quantity of water containing the positron-emitting radioactive isotope oxygen-15 (^{15} O). The radioactive water accumulates in the brain in direct proportion to the local blood flow, so areas to which the most blood is flowing will show the greatest concentration of ^{15} O.

The PET scanner tracks the progress of the radioactive water through the brain (for about a minute, before the radioactive isotope decays to a nonradioactive atom). This provides an indirect but highly reliable measure of blood flow in the brain and, hence, a way of telling which brain regions are active during the minute after administering the water. If subjects are carrying out particular experimental tasks during that time, then the PET technology gives scientists a tool for identifying which brain regions are actively involved in carrying out that task.

However, simply identifying which brain regions have blood flowing to them while a particular task is being performed is not enough to tell us which brain regions are actively involved in carrying out the task. There may be all sorts of activity going on in the brain that is not specific to the particular experiment that the subject is performing. The art in designing PET experiments is finding ways to filter out potentially irrelevant

background activity. The experiments we will be focusing on, carried out by Steve Petersen and collaborators at Washington University in St. Louis, provide a very nice illustration of how this sort of filtering can be done — and of how careful experimental work can refine information-processing models.

Petersen et al., " Positron Emission Tomographic Studies of the Cortical Anatomy of Single-Word Processing " (1988)

Petersen and his colleagues were studying how linguistic information is processed in the human brain. They started with individual words — the basic building blocks of language. Many different types of information are relevant to the normal course of reading, writing, or conversing. There is visual information about the shape and layout of the word, as well as auditory information about how the word sounds and *semantic* information about what the word means. The interesting question is how these different types of information are connected together. Does silently reading a word to oneself involve processing information about how the word sounds? Does simply repeating a word involve recruiting information about what the word means?

The two leading information-processing models of single-word processing (often called *lexical access*) answer these two questions very differently. Within neurology, the dominant model, derived primarily from observing patients with brain damage, holds that the processing of individual words in normal subjects follows a single, largely invariant path. The information-processing channel begins in the sensory areas. Auditory information about how the word sounds is processed in a separate brain

region from information about the word ′s visual appearance. According to the neurological model, however, visual information about the word ′s appearance needs to be phonologically recoded before it can undergo further processing. So, in order to access semantic information about what a written word means, the neurological model holds that the brain needs to work out what the word sounds like. Moreover, in this model, semantic processing is an essential preliminary to producing phonological motor output. So, for example, the act of reading a word and then pronouncing it aloud involves recruiting information about what the word means.

Exercise 3.5 Draw a flowchart illustrating the distinct information-processing stages in single-word processing according to the neurological model.

The principal alternative to the neurological model is the cognitive model (derived primarily from experiments on normal subjects rather than from studies of patients with brain damage). The neurological model is *serial.* It holds that information travels through a fixed series of information-processing " stations " in a fixed order. In contrast, the cognitive model holds that lexical information processing is *parallel.* The brain can carry out different types of lexical information processing at once, with several channels that can feed into semantic processing. Likewise, there is no single route into phonological output processing.

Petersen and colleagues designed a complex experiment to determine which model more accurately reflects the channels of lexical information processing in the brain. The basic idea was to organize the experimental

conditions hierarchically so that each condition could tap into a more advanced level of information processing than its predecessor. Each level involved a new type of information-processing task. Successfully carrying out the new task required successfully carrying out the other tasks lower in the hierarchy. What this means is that by looking at which *new* brain areas are activated in each task, we can identify the brain areas that are specifically involved in performing that task — and we can also see which brain areas are *not* involved.

The baseline condition was simply asking subjects to focus on a fixation point (a small crosshair) in the middle of a television screen. The point of asking the subjects to do this was to identify what is going on in the brain when subjects are visually attending to something that is not a word. The second condition measured brain activity while subjects were passively presented with words flashed on the screen at a rate of forty words per minute. The subjects were not asked to make any response to the words. In a separate condition, the same words were spoken to the subjects.

Combining the results from these two different conditions allowed Petersen and his colleagues to work out which brain areas are involved in visual and auditory word perception. The key to doing this is to subtract the image gained from the first condition from the image derived from the second condition. The image of brain activity while fixating on the crosshair acts as a control state. In principle (and we will look much more closely at some of the methodological difficulties in functional neuroimaging in Chapter 9), this allows us to filter out all the brain activation that is responsible for sensory processing in general, rather than word perception in particular.

The third and fourth levels of the experimental hierarchy traced the connections between initial sensory processing and semantic and output processing further " downstream. " In the third condition, subjects were asked to say out loud the word appearing on the screen. Subtracting the resulting image from the word-perception image allowed Petersen and his colleagues to calculate which brain areas are involved in speech production. In the fourth condition, the subjects were presented with nouns on the television monitor and asked to utter an associated verb (e.g., " turn " for " handlebars "). Subtracting the image of brain activation during this semantic-association task from the image obtained from the speech-production task identified the brain areas involved in semantic processing .

Exercise 3.6 Make a table to show the different levels in the hierarchy and the aspects of single-word processing that they are intended to track.

As we see in Figure 3.9 , each of the tasks activated very different sets of brain areas. (The areas with the maximum blood flow are colored white, followed in decreasing order by shades of red, yellow, green, blue, and purple.)

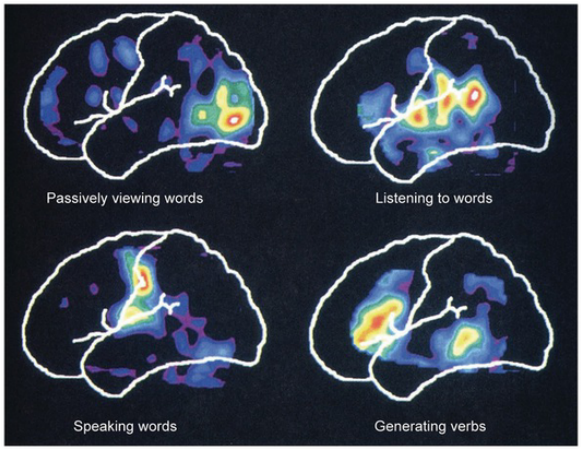


Figure 3.9 Images showing the different areas of activation (as measured by blood flow) during the four different stages in Petersen et al. ' s (1988) lexical access studies.

(From Petersen et al. 1988 .)

These patterns of activation seemed to provide clear evidence against the neurological model because when subjects were asked to repeat visually presented words, there was no activation of the regions associated with auditory processing. This suggested to Petersen and his colleagues that there is a direct information pathway from the areas in the visual cortex associated with visual word processing to the distributed network of areas responsible for articulatory coding and motor programming, coupled with a parallel and equally direct pathway from the areas associated with auditory word processing. Moreover, the areas associated with semantic processing (those identified in the condition at the top of the hierarchy) were not involved in any of the other tasks, suggesting that those direct pathways did not proceed via the semantic areas.

The information-processing diagram in Figure 3.10 comes from a paper by Petersen and collaborators published in *Nature* in 1988. Unlike

many information-processing flowcharts, this one is distinctive in that it identifies the particular brain areas that are thought to carry out each distinct stage. This is not an accident. It reflects how the information-processing model was reached — on the basis of direct study of the brain through PET scanning technology.

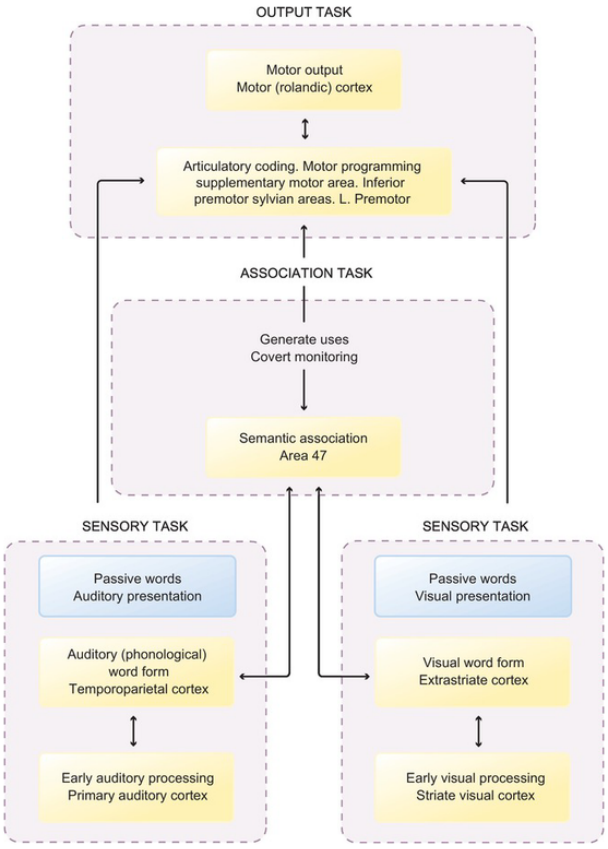


Figure 3.10 A flowchart relating areas of activation in Petersen et al. ' s (1988) study to different levels of lexical processing. The dashed boxes outline the different subtractions. The solid boxes outline possible levels of coding and associated anatomical areas of activation.

(From Petersen et al. 1988 .)

3.5 Studying Memory for Visual Events

There are two principal technologies in functional neuroimaging. Section 3.4 looked at the PET technology, which measures cerebral blood flow by tracking the movement of radioactive water in the brain. A newer, and by now dominant, technology is fMRI.

Functional Neuroimaging with fMRI

The standard background assumption in fMRI (as with PET) is that blood flow to a particular region of the brain increases when cellular activity in that region increases. This increase in blood flow produces an increase in oxygen. Because the supply of oxygen is greater than the demand, the blood oxygen level increases in a brain region that is undergoing increased cellular activity. An fMRI scanner creates a powerful magnetic field, which can detect increases in blood oxygen because oxygenated and deoxygenated blood have different magnetic properties. The difference between oxygenated and deoxygenated blood is known as the *BOLD contrast* . fMRI imaging measures the BOLD signal.

Early fMRI experiments used a similar experimental design to those used in PET studies, such as the Peterson et al. experiments described in Section 3.4 . This is known as a *blocked design* . In a blocked design, experiment subjects perform a task for an extended period of time. The task might be focusing on a fixation point or repeated visually presented words. Using extended blocks maximizes the changes in blood oxygen level and so gives a stronger signal. This makes it easier to compare and contrast signals in different conditions using subtraction methods.

One reason for the popularity of blocked designs in early fMRI experiments is that the BOLD hemodynamic response is delayed in its onset and takes a while to develop. So, for example, for a neural event lasting one second, it takes two seconds for the hemodynamic response to start to develop, and the development takes ten to twelve seconds. So, early researchers concluded that these aspects of the BOLD signal required fMRI experiments to use the same kind of experimental design as PET experiments.

In the early 1990s, however, a new type of experimental design emerged for fMRI experiments. This is usually termed *event-related fMRI.* It is, in some respects, similar to the experimental design used in many electroencephalography (EEG) experiments, which we will look at in more detail in Chapter 9 . Event-related fMRI measures the BOLD signal associated with individual rapidly occurring neural events, even though these events elicit overlapping hemodynamic responses. It is possible because the hemodynamic response that fMRI measures behaves (to a first approximation) like a linear system.

What this means is that the hemodynamic response for a given event in a series basically adds on proportionally to the hemodynamic response for earlier events in the series. So, if you can measure the change in the BOLD signal associated with six task events, each two seconds in length, the result will be close to the BOLD signal associated with a single task event, lasting for twelve seconds. Because of this linearity, even though the hemodynamic responses for those six events are all overlapping, standard statistical techniques can be used to work backward from the overall change in the BOLD signal for the series to the particular change in the BOLD signal for

each event. And it is also possible, although more complicated, to do this when the events in the series are different from each other.

With the use of event-related fMRI, neuroscientists are able to study the BOLD signal produced by short-duration events, and they are also able to disentangle the separate components of complex tasks. This is a very significant step beyond the blocked-design paradigm (although that remains useful for various areas of neuroimaging, such as neuroimaging for language). To illustrate the power of event-related fMRI, we will look at an important early study on memory for visual events.

Brewer et al., " Making Memories: Brain Activity That Predicts How Well Visual Experience Will Be Remembered " (1998)

This paper was one of the first fMRI studies to use an event-related experimental design. It was trying to get at something that could not be uncovered using a blocked design because it depended on information about changes in the BOLD signal brought about by very specific and short-lived neural events. Brewer and his colleagues were interested in exploring whether there are any neural markers predicting how well specific visual experiences would be remembered. Are there any areas in the brain where activity in those areas would predict whether they would be remembered well, remembered less well, or forgotten?

It is important to realize what this question is *not* asking. It is not asking which areas in the brain are involved in memory. That was already fairly well known from studies of patients with brain damage. So, for example, there is considerable evidence that damage to the medial temporal lobe causes global amnesia and that damage to specific parts of the frontal

lobes can bring about different types of local amnesia. Damage to the left frontal lobes affects verbal memory, for example.

But looking at brain damage cannot help us distinguish between the different aspects of memory. It cannot tell us whether what is impaired is how experiences are encoded for memory, how memories are stored, or how they are retrieved. And it certainly cannot tell us anything about individual experiences and how likely they are to be remembered. For that, you need to be able to identify the specific hemodynamic response generated by specific experiences — in other words, you need an event-related design.

Exercise 3.7 Explain in your own words the difference between a blocked design and an event-related design.

To get at this question about how individual experiences are encoded in memory, the experimenters showed subjects in an fMRI scanner ninety-six color pictures of indoor and outdoor scenes over four trials. The pictures were selected to be broadly comparable in complexity and visual quality so that they would all make roughly the same general processing demands. All that the subjects were asked to do while in the scanner was to identify, for each picture, whether it was an indoor or outdoor scene. Then, thirty minutes later, the subjects were given an unanticipated memory test, presented with 128 pictures, including the 96 shown to them in the scanner, and asked to identify which they had seen before and to identify how confident they were.

The memory test allowed the experimenters to classify how well each subject remembered each of the originally displayed pictures — as well remembered, familiar, or forgotten. The next step was to try to find patterns of neural activity correlated with each of those three levels of memory performance. To that end, the experimenters constructed two maps of event-related activity. The first map showed brain areas where the hemodynamic response increased when subjects were looking at pictures (relative to activation levels during fixation). So, this map measured the BOLD response generated by visual experiences. The first map is illustrated in Figure 3.11 .

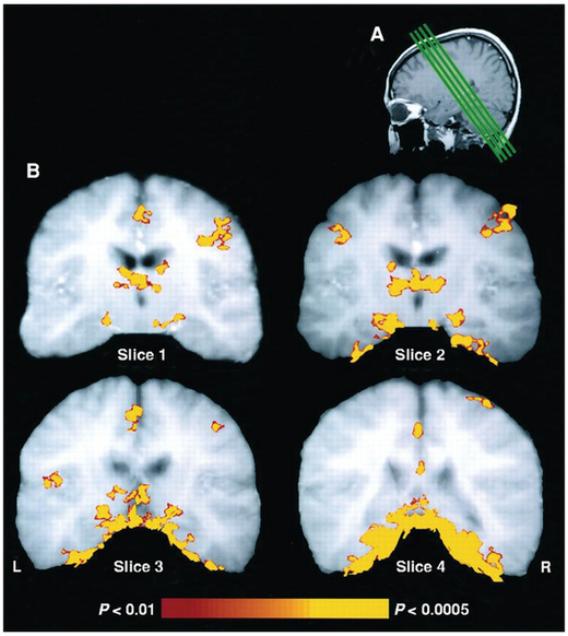


Figure 3.11 Neural areas showing activity when subjects looked at pictures.

(From Figure 1 in Brewer et al. 1998 .)

As Figure 3.11 shows, many areas are involved in visual experience. But which ones are responsible for encoding visual experiences into

memory?

The second map, depicted in Figure 3.12 , answered this question. It showed brain areas where activation levels were correlated with the levels of memory performance for individual events. So, for example, a brain area would appear on the second map if it showed high levels of increased activation during visual experiences of pictures that were subsequently well remembered, medium levels of increased activation during visual experiences that were subsequently judged familiar, and low levels of increased activation during visual experiences that were not remembered.

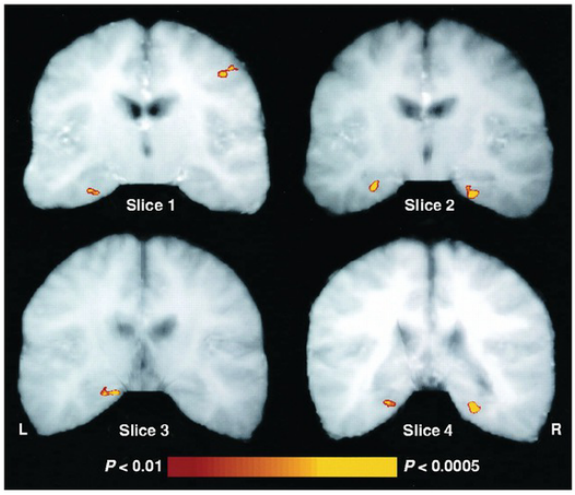


Figure 3.12 Neural areas where activation is correlated with levels of memory performance.

(From Figure 2 in Brewer et al. 1998 .)

Comparing Figures 3.11 and 3.12 shows that encoding visual experiences into memory is highly localized. Of the many areas implicated in visual experience, only two predict how well those visual experiences will be remembered. The first area is the parahippocampal cortex (in both hemispheres). The parahippocampal cortex is part of the medial temporal

lobe. The second area is the dorsolateral prefrontal cortex, only in the right hemisphere .

Quite apart from the intrinsic interest of these results for the study of memory, they are an excellent illustration of the power of event-related designs for fMRI. The event-related design makes it possible to identify the hemodynamic response generated by each individual visual experience, and without that, it would be impossible to identify the areas that predict which experiences will be remembered — as opposed to identifying the areas generally responsible for visual experience and/or memory.

3.6 The Neural Correlates of the BOLD Signal

When one is looking at brightly colored pictures communicating the results of PET or fMRI scans, it is only too easy to forget that relatively little is known about the relation between what those scans measure and the cognitive activity that is going on while the measurements are being made. It is only in the very recent past that progress has been made on building a bridge between functional neuroimaging and neurophysiology.

As we saw in Section 3.5 , fMRI measures the BOLD contrast. But what does the BOLD contrast measure? In some sense, the BOLD contrast has to be an index of cognitive activity — because it is known that cognitive activity involves increased activity in populations of neurons, which in turn results in increased oxygen levels and, hence, in a more pronounced BOLD contrast. But what sort of neuronal activity is it that generates the BOLD contrast?

Neuroscientists study the behavior of individual neurons through single-cell recordings (to be discussed in more detail in Chapter 9). Microelectrodes can be inserted into the brains of animals (and also of humans undergoing surgery) and then used to record activity in individual cells while the animal performs various behavioral tasks. Figure 3.13 illustrates a microelectrode recording in the vicinity of a single neuron.



Figure 3.13 A microelectrode making an extracellular recording.

(Scientific American Library [W. H. Freeman 1995].)

Because the salient property of individual neurons is their firing (or *spiking*) behavior, it is a natural assumption that the neural activity correlated with the BOLD contrast is a function of the firing rates of populations of neurons. In fact, this is exactly what was suggested by Geraint Rees, Karl Friston, and Christoph Koch in a paper published in 2000. They proposed that there is a linear relationship between the average

neuronal firing rate and the strength of the BOLD signal — two variables are linearly related when they increase in direct proportion to each other, so if one were to plot their relation on a graph, it would be a straight line.

This conclusion was based on comparing human fMRI data with single-cell recordings from monkeys. In fact, their study seemed to show that each percentage increase in the BOLD contrast is correlated with an average per-second increase of nine spikes per unit. If the Rees – Friston – Koch hypothesis is correct, then the BOLD response directly reflects the average firing rate of neurons in the relevant brain area, so an increase in the BOLD contrast is an index of higher neural firing activity.

Neurons do more than simply fire, however. We can think of a neuron ′ s firing as its *output.* When a neuron fires, it sends a signal to the other neurons to which it is connected. This signal is the result of processing internal to the neuron. This processing does not always result in the neuron ′ s firing. Neurons are selective. They fire only when the level of internal activity reaches a particular threshold. This means that there can be plenty of activity in a neuron even when that neuron does not fire.

We might think of this as a function of the *input* to a neuron rather than of its output. A natural question to ask, therefore, is whether the BOLD signal is correlated with the input to neurons or with their output (as Rees and colleagues had proposed). This is exactly the question explored in a very influential experiment by Nikos Logothetis and collaborators.

Logothetis, " The Underpinnings of the BOLD Functional Magnetic Resonance Imaging Signal " (2001)

Logothetis compared the strength of the BOLD signal against different measures of neural activity in the monkey primary visual cortex (see Section 3.2 for a refresher on where the primary visual cortex is and what it does). The team measured neural activity in an anesthetized monkey when it was stimulated with a rotating checkerboard pattern while in a scanner. In addition to using fMRI to measure the BOLD contrast, researchers used microelectrodes to measure both input neural activity and output neural activity. This is particularly challenging from an engineering point of view. Because an fMRI scanner generates a powerful magnetic field, the microelectrodes need to be nonmagnetic.

At the output level, they measured the firing rates of both single neurons and small populations of neurons near the electrode tip (" near " here means within 0.2 mm or so). In Figure 3.14 , these are labeled *SDF* (spike density function) and *MUA* (multiunit activity).

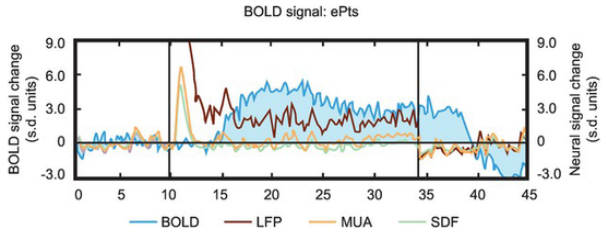


Figure 3.14 Simultaneous microelectrode and fMRI recordings from a cortical site showing the neural response to a pulse stimulus of twenty-four seconds. Both single- and multiunit responses adapt a couple of seconds after stimulus onset, with the local field potential remaining the only signal correlated with the BOLD response.

(Adapted from Bandettini and Ungerleider 2001 .)

The *local field potential* (LFP) is an electrophysiological signal believed to be correlated with the sum of inputs to neurons in a particular area. It is also measured through a microelectrode, but the signal is passed through a *low-pass filter* that smooths out the quick fluctuations in the signal that are due to neurons firing and leaves only the low-frequency signal that represents the inputs into the area to which the electrode is sensitive (an area a few millimeters across).

Logothetis and his team concluded that the BOLD contrast is more highly correlated with the LFP than with the firing activity of neurons (either at the single-unit or multiunit level). This is nicely illustrated in the graph in Figure 3.14 . In many cases, the LFP will itself be correlated with the firing activity of neurons (which is why Logothetis ' s results are perfectly compatible with the results reached by Rees and colleagues). But, if Logothetis ' s data do indeed generalize, then they show that when spiking activity and LFP are *not* correlated, the LFP is the more relevant of the two to the BOLD contrast .

Summary

This chapter illustrated the " turn to the brain " that took place in cognitive science during the 1980s. The first example we looked at was the two-visual-systems hypothesis developed by Mishkin and Ungerleider, primarily on the basis of monkey experiments. We then looked at the emergence of parallel-distributed-processing models in the 1980s and reviewed a famous application by Gorman and Sejnowski to the problem of distinguishing rocks from mines in sonar recordings. The last three sections focused on

neuroimaging. We began with PET and reviewed an influential set of experiments on single-word processing from Petersen et al. For many applications, PET has been superseded by fMRI, not least because it allows event-related experimental designs, as opposed to the blocked designs used in PET and early fMRI studies. We then looked at the power of the event-related design in the context of experiments by Brewer and colleagues on predicting how well visual experiences would be remembered. Finally, we turned to the neural correlates of the BOLD contrast, which is what fMRI measures directly. The BOLD contrast is a hemodynamic response, a function of blood oxygen levels. But what do blood oxygen levels tell us about neural activity? An elegant set of experiments by Nikos Logothetis tackles this question.

Checklists

Ungerleider and Mishkin ' s Two-Visual-Systems Hypothesis

1. The cross-lesion disconnection paradigm, coupled with various other anatomical and neurological methods, was used to identify two different information-processing pathways for visual information.

2. Both pathways start from the primary visual cortex.

3. Information relevant to object identification and recognition travels along the ventral pathway, from the primary visual cortex to the inferior temporal cortex via the prestriate cortex.

4. Information relevant to locating objects flows from the primary visual cortex to the posterior parietal lobe.

Information Processing in Artificial Neural Networks

1. These networks are designed to reflect high-level features of how the brain processes information, such as its parallel and distributed nature.

2. The neuronlike units in artificial neural networks are organized into layers, with no connections between the units in a single layer.

3. The overall behavior of the network is determined by the weights attached to the connections between pairs of units in adjacent layers.

4. Networks " learn " by adjusting the weights in order to reduce error.

5. Artificial neural networks are particularly suited to pattern-recognition tasks, such as discriminating sonar echoes caused by mines from those caused by rocks.

Functional Neuroimaging: PET and the Example of Single-Word Processing

1. PET allows brain activity to be studied noninvasively by measuring blood flow in the brain while subjects are performing particular cognitive tasks.

2. The paired-subtraction paradigm focuses on the brain activity specific to the task by subtracting out the activity generated by carefully chosen control tasks.

3. In studies of how single words are processed, experimenters constructed a four-level hierarchy of tasks of increasing complexity.

4. The patterns of activation they identified across the different tasks supported a parallel rather than a serial model of single-word processing.

Functional Neuroimaging: Event-Related fMRI and Predicting How Well Visual Experiences Will Be Remembered

1. Functional magnetic resonance imaging (fMRI) measures levels of blood oxygen in the brain (the blood-oxygen-level – dependent [BOLD] contrast/signal) and has superseded positron emission tomography (PET) for many applications.

2. Event-related fMRI allows researchers to study the BOLD signal associated with individual neural events, unlike the blocked-design standard used in PET imaging.

3. Event-related fMRI works because even though the BOLD response is delayed and takes time to develop, it behaves linearly, which allows a cumulative BOLD signal derived from multiple individual events to be broken down into its constituent elements.

4. Event-related fMRI was illustrated through a study of how individual visual experiences are encoded into memory.

Neural Correlates of the BOLD Signal

1. fMRI provides a measure of blood flow in terms of levels of blood oxygenation (the BOLD signal), giving an index of cognitive activity.

2. This raises the question of how this cognitive activity is related to neural activity.

3. One possibility is that the cognitive activity detected by fMRI is correlated with the outputs of populations of neurons (as manifested in their firing activity). Another possibility is that the correlation is with the input to populations of neurons (as measured by the local field potential).

4. The experiments of Logothetis and his collaborators suggest that the correlation is with the input to neural areas rather than with their output.

Further Reading

Ungerleider and Mishkin ' s paper " Two Cortical Visual Systems " is reprinted in Cummins and Cummins 2000 . Mishkin et al. 1983/2001 is a little more accessible. David Milner and Melvyn Goodale have developed a different version of the two-visual-systems hypothesis, placing much more emphasis on studies of patients with brain damage. See, for example, their book *The Visual Brain in Action* (Milner and Goodale 2006). A more recent summary can be found in Milner and Goodale 2008 (including discussion of Ungerleider and Mishkin). A different development in terms of vision for action versus vision for higher mental processes has been proposed by cognitive neuroscientist Marc Jeannerod, as presented in *Ways of Seeing* , coauthored with the philosopher Pierre Jacob (Jacob and Jeannerod 2003). A critique of the two-system account (with commentary from Milner, Goodale, and others) can be found in Schenk and McIntosh 2010 . See Rossetti et al. 2017 for another critical perspective overview.

The Handbook of Brain Theory and Neural Networks (Arbib 2003) remains the most comprehensive single-volume source for different types of computational neuroscience and neural computing, together with entries on neuroanatomy and many other " neural topics. " It contains useful introductory material and " road maps. " Dayan and Abbott 2005 and Trappenberg 2010 are other commonly used introductory textbooks. Scholarpedia.org is also a good source for introductory articles specifically on topics in computational neuroscience. McLeod et al. 1998 is an introduction to connectionism that comes with software allowing readers to get hands-on experience in connectionist modeling. Bechtel and Abrahamsen 2002 is also to be recommended. Useful article-length presentations are Rumelhart 1989 (in Posner 1989 , reprinted in Haugeland 1997) and Churchland 1990b (in Cummins and Cummins 2000). A more recent discussion of connectionism can be found in McClelland et al. 2010 , with commentary and target articles from others in the same issue. The mine/rock network described in the text was first presented in Gorman and Sejnowski 1988 and is discussed in Churchland 1990a . There are more references to literature on neural networks in the Further Reading sections for Chapters 5 and 10 . Recent interest in neural networks has been associated with deep learning, discussed in Chapter 12 .

A very readable book introducing PET and functional neuroimaging in general is Posner and Raichle 1994 , written by two senior scientists participating in the lexical-access experiments discussed in the text. These experiments are discussed in the article by Petersen et al. cited in the text and also (more accessibly) in Petersen and Fiez 2001 . Fiez et al. 1999 explores similar issues using fMRI rather than PET. Rowe and Frackowiak 2003 is an article-length introduction to the basic principles of functional

neuroimaging. Another good introduction to neuroimaging, including discussion of many of the experiments mentioned in this chapter (and with a lot of colorful illustrations), is Baars and Gage 2010 .

For the specifics of event-related fMRI, see the overviews in Buckner 1998 and Huettel 2012 . The Huettel paper is published in a special issue of the journal *Neuroimage* edited by Peter Bendettini and titled " 20 Years of fMRI " (62:2, August 2012). For the first study using event-related fMRI, see Blamire et al. 1992 . Dale and Buckner 1997 is another early study. The study referenced in the text is Brewer et al. 1998 . See also Wagner et al. 1998 in the same issue of *Science.* Xue 2018 gives an overview of neural representations in episodic memory.

For specific references on fMRI technology, see the Further Reading suggestions in Chapter 9 . For a survey of some of the general issues in thinking about the neural correlates of the BOLD signal, see Heeger and Ress 2002 and Raichle and Mintun 2006 . Logothetis ' s single-authored 2001 paper in the *Journal of Neuroscience* is a good introduction to the general issues as well as to his own experiments. See also Logothetis 2002 . A more recent summary can be found in Goense et al. 2012 . For the Rees – Friston – Koch hypothesis, see Rees et al. 2000 . For commentary on Logothetis, see Bandettini and Ungerleider 2001 . For an alternative view, see Mukamel et al. 2005 . Ekstrom 2010 discusses apparent dissociations between the BOLD signal and local field potentials, whereas Zhang et al. 2019 suggest that the relation may be nonlinear.

Part II



Models and Tools

Chapter 4

Physical Symbol Systems and the Language of Thought



Overview

4.1 The Physical Symbol System Hypothesis

Symbols and Symbol Systems

Transforming Symbol Structures

Intelligent Action and the Physical Symbol System

4.2 From Physical Symbol Systems to the Language of Thought

Intentional Realism and Causation by Content

The Language of Thought and the Relation between Syntax and Semantics

4.3 The Russian Room Argument and the Turing Test

Responding to the Russian Room Argument

Overview

The analogy between minds and digital computers is one of the most powerful ideas in cognitive science. The physical symbol system hypothesis, proposed in 1975 by computer scientists Herbert Simon and Allen Newell, articulates the analogy very clearly. It holds that all intelligent behavior essentially involves transforming physical symbols according to rules. Section 4.1 explains the basic idea, and Section 4.2 looks at the version of the physical symbol system hypothesis developed by philosopher Jerry Fodor. Fodor develops a subtle and sophisticated argument for why symbolic information processing has to take place in a language of thought.

Both hypotheses rest on a sharp distinction between the syntax of information processing (the physical manipulation of symbol structures) and the semantics of information processing (the meanings of those symbols). Philosopher John Searle has developed a famous argument (originally called the *Chinese room argument*) aiming to show that the project of modeling the mind as a computer is fatally flawed because syntax can never yield semantics. We look at a version of his argument and at some of the ways of replying to it in Section 4.3 .

4.1 The Physical Symbol System Hypothesis

In 1975 the Association of Computing Machinery gave its annual Turing Award to two very influential pioneers of artificial intelligence (AI) — Herbert Simon and Allen Newell. Simon was actually an economist and political scientist rather than a computer scientist (as Newell was). Their

joint contributions to cognitive science included the Logic Theory Machine (1957) and the General Problem Solver (1956), two early and very important programs that developed general strategies for solving formalized symbolic problems.

Newell and Simon gave a public lecture as one of the conditions of receiving the award. That lecture proposed a bold strategy both for the study of the human mind and for the emerging field of AI. It hinged on what they called the *physical symbol system hypothesis.*

For Newell and Simon, the physical symbol system hypothesis is as basic to AI as the principle that the cell is the basic building block of all living organisms is to biology. Here is how they phrased it:

> ***The physical symbol system hypothesis*** **:** A physical symbol system has the necessary and sufficient means for general intelligent action.

There are two separate claims here. The first (the necessity claim) is that nothing can be capable of intelligent action unless it is a physical symbol system. So, because humans are capable of intelligent action, the human mind must be a physical symbol system. The second (the sufficiency claim) is that in principle, there is no obstacle to constructing an artificial mind, provided that one tackles the problem by constructing a physical symbol system.

For Newell and Simon, physical symbol systems have four distinctive features:

1. Symbols are physical patterns.

2. These symbols can be combined to form complex symbol structures.

3. The physical symbol system contains processes for manipulating symbols and complex symbol structures.

4. The processes for generating and transforming complex symbol structures can themselves be represented by symbols and symbol structures within the system.

As you might have noticed, a physical symbol system looks very much like an abstract characterization of a digital computer.

Symbols and Symbol Systems

Turing machines, which we first encountered in Section 1.2 , nicely illustrate the first two ideas in the physical symbol system hypothesis:

1. *Symbols are physical patterns.* For Newell and Simon, symbols are physical objects, just as written letters on a page are physical objects, as are spoken words (which are soundwaves). The symbols in Turing machines are also physical objects. They are inscriptions on the tape that the Turing machine is able to read.

2. *Symbols can be combined to form complex symbol structures.* Just as letters can be put together to form words, the symbols in any physical symbol system can be combined to form word-like symbol structures. Those word-like structures can then be put together to form sentence-like structures. Both types of combinations are governed by strict rules. You can think of these strict rules as telling the symbol system which combinations of symbols count as grammatical.

These rules are typically *recursive* . That means that they will show how to get from an acceptable combination of symbols to a more complex

combination that is still acceptable. The rules for how to define what counts as a sentence in the branch of logic known as *sentential logic* or *propositional logic* provide a good illustration of recursive rules and how they work. See Box 4.1 .

BOX 4.1 Defining Sentences in Propositional Logic

Propositional logic studies the logical relations holding between whole sentences, or propositions. The language of propositional logic is very simple. It contains basic symbols for sentences (such as " P, " " Q, " and " R "), together with a small set of logical connectives.

A typical formulation of propositional logic might have three connectives (the so-called *Boolean connectives*). These are " ¬, " read as " not " ; " ∨ , " read as " or " ; and " ∧ , " read as " and. "

These logical connectives allow sentence symbols to be combined to form more complex sentences. So, for example, " P ∧ Q " is a sentence. It is true only when the two sentences P and Q are both true.

Propositional logic has clear and unambiguous rules for determining what counts as a legitimate sentence. These rules determine when the rules governing the connectives have been correctly applied. The legitimate combinations of symbols in the alphabet might typically be defined as follows:

a. Any sentence symbol is a sentence.

b. If " φ " is a sentence, then " ¬ φ " is a sentence.

c. If " φ " and " ψ " are sentences, then " φ ∧ ψ " is a sentence.

d. If " φ " and " ψ " are sentences, then " φ ∨ ψ " is a sentence.

These are examples of *recursive rules.* They show how, starting with a basic set of sentences (the sentence symbols), you can construct

arbitrarily complex formulas that will still be genuine sentences.

Note that " φ " and " ψ " can stand here for any formula, not just for sentence symbols. So, you can apply the recursive definition to show that ¬(P ∧ ¬P) is a genuine sentence of propositional logic.

Can you see how? (Hint: If P is a sentence symbol, then it is a sentence, by item a in the previous list. If P is a sentence, then so is ¬P, by item b. Continue in this vein.)

Turing machines can scan only a single cell at a time, but they are still capable of working with complex symbol structures because those complex symbol structures can be built up from individual symbols in adjacent cells. The Turing machine needs to know what symbols can follow other symbols, and it needs some way of marking the end of complex symbols. The first can come from instructions in the machine table, whereas for the second, there are symbols that serve as punctuation marks, effectively telling the scanner when it has arrived at the end of a complex symbol.

Transforming Symbol Structures

The third feature of physical symbol systems involves transformation:

3. *The physical symbol system contains processes for manipulating symbols and complex symbol structures* . This is the distinctive claim of the physical symbol system hypothesis. Thinking is no more (and no less) than transforming symbol structures according to rules. According to Newell and Simon, when we fully understand what is going on in intelligent agents (such as human beings), all we will ultimately find is symbol structures being transformed in rule-governed ways.

In the background here is Newell and Simon ' s fundamental idea that the essence of intelligent thinking is the ability to solve problems. Intelligence consists of the ability to work out, when confronted with a range of options, which of those options best matches certain requirements and constraints.

Intelligence cannot be applied without what might abstractly be called a *search-space* . The notion of a search-space is very general. Consider, for example, the position of one of the players halfway through a chess match. Each chess player has a large number of possible moves and a clearly defined aim — to checkmate the opponent. The possible moves define the search-space, and the problem is deciding which of the possible moves will move the player closest to their goal.

Another example (much studied by computer scientists and mathematicians) is a traveling salesperson who starts in a particular city (say, Boston) and has to visit twenty other cities as quickly and efficiently as possible before eventually returning to Boston. Here, we can think about the search-space in terms of all the possible routes that start and end in Boston and go through the twenty cities (perhaps visiting some more than once). The diagram at the top in Figure 4.1 illustrates a simpler traveling salesperson problem with only five cities (*a* , *b* , *c* , *d* , and *e*).

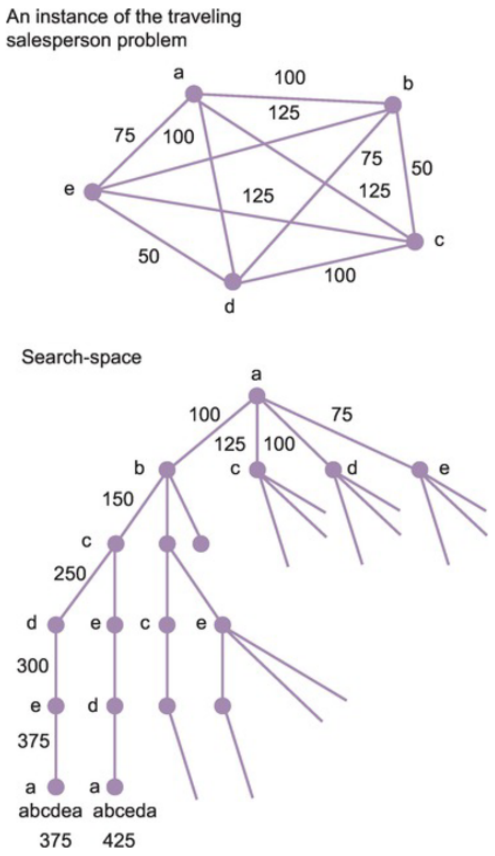


Figure 4.1 A typical traveling salesperson problem. The top diagram depicts the problem. A traveling salesperson has to find the shortest route between five cities. The bottom diagram depicts part of the search-space. A complete representation of the search-space would show twenty-four different routes.

(From Figures 3-7 and 3-8 in Luger 2001 .)

Search-spaces are typically represented in terms of states. There is an initial state (the start state) and a set of permissible transformations of that start state. The search-space consists of all the states that can be reached from the start state by applying the permissible transformations. The transformations can be carried out in any order. In the chess example, the start state is a particular configuration of the chess pieces, and the permissible transformations are the legal moves in chess. In the traveling

salesman problem, the start state might be Boston, for example, and the permissible transformations are given by all the ways of getting directly from one city to another.

Computer scientists standardly represent search-spaces in terms of trees. So, for example, the search-space for the traveling salesperson problem is given by a tree whose first node is the starting city. The diagram at the bottom of Figure 4.1 illustrates a part of the search-space for our five-city version of the traveling salesperson problem. A branch from the first node (*a* , the start city) goes to a node representing each city to which the start city is directly connected — that is, cities *b* , *c* , *d* , and *e* . From each of those nodes, further branches connect each city to all the other cities to which it is directly connected. And so on.

What counts as solving a problem? Basically, you ′ ve solved a problem when you ′ ve found the solution state in the search-space. In chess, the solution state is any configuration of the board in which the opponent ′ s king is in checkmate. In the traveling salesperson problem, the solution is the shortest branch of the tree that ends with Boston and that has nodes on it corresponding to each city that the salesperson needs to visit.

But how is that done? Obviously, you have to search through the search-space until you find a solution state. But this can be much harder than it sounds. Brute-force searches that follow each branch of the tree typically only work for very simple problems. It does not take long for a problem space to get so big that it cannot be exhaustively searched in any feasible amount of time.

The traveling salesperson problem is a great example. If there are *n* cities, then it turns out that there are (*n* − 1)! possible routes to take into account, where $(n-1)! = (n-1) \times (n-2) \times (n-3) \ldots$ This number of

routes is not too many for the five-city version of the problem depicted in Figure 4.1 (it gives twenty-four different routes). But the problem gets out of control very quickly. A twenty-city version gives approximately 6×10^{16} different ways for a traveling salesperson to start in Boston and travel through the other nineteen cities, visiting each exactly once. Checking one route per second, we would need more than the entire history of the universe to search the problem space exhaustively.

Newell and Simon developed their General Problem Solver (GPS) program as a way of solving problems of this type (although they focused on much simpler problems than the traveling salesperson problem, which still has no general solution).

The basic idea behind the GPS program is relatively straightforward. The program uses *means – end analysis*. Here are simple instructions for applying means – end analysis:

1. Evaluate the difference between the current state and the solution state.

2. Identify a transformation that reduces the difference between the current state and the solution state.

3. Check that the transformation in step 2 can be applied to the current state.

3a. If it can, then apply it and go back to step 1.

3b. If it can ′ t, then return to step 2.

Means – end analysis is an example of what Newell and Simon call *heuristic search.* Heuristic search techniques are techniques for searching through a search-space that do not involve exhaustively tracing every branch in the

tree until a solution is found. Heuristic search techniques reduce the size of the search-space in order to make the search process more manageable.

Exercise 4.1 Explain how means – end analysis makes it more manageable to search the search-space.

Here is the problem of the foxes and the chickens — a type of problem that Newell and Simon showed could be solved by their GPS program. Imagine that there are three chickens and three foxes on one side of a river, and they all need to get over to the other side. The only way to cross the river is in a boat that can take only two animals (or fewer) at a time. The boat can cross in either direction, but if at any point the foxes outnumber the chickens (either onshore or in the boat), they will eat the outnumbered chickens. How can you get all the chickens and foxes onto the other side of the river without any of the chickens being eaten?

Here, each state specifies which animals are on each bank and which are in the boat (as well as the direction in which the boat is traveling). The start state obviously has all six on one bank (say, the left bank), with nobody in the boat or on the other bank. The solution state is the state that has all six on the right bank, with nobody in the boat or on the other bank. The permissible transformations are defined by the rule that the boat cannot carry more than two animals at a time.

The foxes and chickens problem is a great example of how the GPS program works. If we feed representations of the start state and the solution, or goal state, into the GPS program, the program employs various transformation strategies to minimize the difference between the start state

and the goal state. The eventual solution is a series of representations, whose first member is a representation of the start state and whose final member is a representation of one of the goal states, where each member is derived from its predecessor by a permissible transformation.

Exercise 4.2 Find a solution to the foxes and chickens problem.

A final point. These rule-governed transformations are *algorithmic.* An algorithm is a finite set of unambiguous rules that can be applied systematically to transform an object or set of objects in definite and circumscribed ways. Algorithms are purely mechanical procedures. They can be followed blindly, without any exercise of judgment or intuition. Elementary school arithmetic provides plenty of examples of algorithms, such as the algorithms for multiplying pairs of numbers and for long division.

Intelligent Action and the Physical Symbol System

We turn to the final feature of physical symbols:

4. *The processes for generating and transforming complex symbol structures can themselves be represented by symbols and symbol structures within the system.* A fundamental feature of modern computers — so familiar that most of us never think about it — is the fact that a single computer (a single piece of hardware) can run many different programs, often simultaneously. This capability is what distinguishes a general-purpose computer from a specialized computing machine such as a pocket calculator. Computers can be programmed in

this way because they can contain symbol structures that encode information about, and instructions for, other symbol structures.

Alan Turing proved that it is possible to construct a special kind of Turing machine (a *universal Turing machine*) that can mimic any specialized Turing machine implementing a particular algorithm. The universal Turing machine is a general-purpose computer. You can think of the specialized computers as software programs that run on the more general operating system of the universal Turing machine. The universal Turing machine is possible because Turing machine tables can be encoded as numbers and, hence, can serve as inputs to Turing machines. The physical symbol system hypothesis builds something like this feature into the characterization of an intelligent system .

4.2 From Physical Symbol Systems to the Language of Thought

The physical symbol system hypothesis tells us that intelligent agents solve problems by physically transforming symbolic structures. But we still need to know what these symbolic structures are, how they are transformed, and how those transformations give rise to intelligent action of the sort that human beings might carry out. This section looks at a proposal for answering these questions. This is the *language of thought hypothesis* developed by philosopher and cognitive scientist Jerry Fodor (1935－2017).

According to Fodor ' s language of thought hypothesis, the basic symbol structures in the mind that carry information are sentences in an internal language of thought (sometimes called *Mentalese* and generally

abbreviated as *LoT*). Information processing works by transforming those sentences in LoT.

Our starting point for exploring this idea is the basic fact that the mind *receives* information about its environment. Some of this information is carried by light waves arriving at the retina or sound waves hitting the eardrum. But in general, our behavior is not *determined* by the information that we receive. Different people, or the same person at different times, will react differently to the same situation. There is no standard response to the pattern of sound waves associated (in English) with, for example, a cry of " Help! " How we behave depends on what our minds do with the information that they receive — how they *process* that information. If I run to your assistance when you cry " Help, " it is because my mind has somehow managed to decode your utterance as a word in English, worked out what you are trying to communicate, and then decided how to respond. This is all complex processing of the initial information that arrived at my eardrum.

But how does this information processing take place? The basic idea of the physical symbol system is that there are physical structures that carry information and, by so doing, serve as *representations* of the immediate environment (or, of course, of things that are more abstract and/or more remote).

Information processing involves many different kinds of representation. This is illustrated by the example just given. The whole process begins with representations that carry information about vibrations in the eardrum. Somehow these representations get transformed into my *belief* that you are in danger. This belief is the " motor " of my behavior (my jumping into the water to rescue you). But it is not enough on its own. It needs to interact with other representations (such as my belief that I can reach you before you

drown and my *desire* to rescue you) in order to generate an *intention* to act in a certain way. This intention, in turn, gives rise to further representations, corresponding to the motor instructions that generate and control my bodily movements.

Among all these different types of representation, Fodor is particularly interested in those that correspond to beliefs, desires, and other similar psychological states. These psychological states are often called *propositional attitudes* by philosophers. They are called this because they can be analyzed as attitudes toward propositions. Propositions are the sorts of things that are expressed by ordinary sentences. So, there is a proposition expressed by the sentence " That person will drown " or by the sentence " It is snowing in St. Louis. " Thinkers can have different attitudes toward those propositions. I might fear the first, for example, and believe the second.

Fodor ' s starting point in thinking about propositional attitudes is that we are, by and large, pretty good at explaining and predicting other people ' s behavior in terms of what they believe about the world and what they want to achieve. He thinks that this success is something that itself needs explanation. Why is the vocabulary of beliefs and desires (our *belief - desire psychology* or *propositional attitude psychology*) so deeply ingrained in us? Why does it seem so indispensable in our social interactions and social coordination? How and why do explanations that appeal to beliefs and desires actually work? And in particular, why are these explanations so successful?

According to Fodor, there can be only one possible answer. Belief - desire psychology is successful because it is true. There really are such things as beliefs and desires. They are physical items that cause us to behave in certain ways. Belief - desire explanations are successful when they

correctly identify the beliefs and other states that caused us to act in the way that we did.

If we say that someone jumped into the water because she believed that a child was drowning and wanted to save him, then what we are really claiming is that that person ' s bodily behavior was caused by internal items corresponding to the belief that someone is drowning and the desire to save her. This view is often called *intentional realism.*

Fodor ' s argument for the LoT hypothesis is, in essence, that intentional realism is the only way of explaining how belief – desire explanations can work. We will examine his argument in the next two subsections.

Exercise 4.3 Explain intentional realism in your own words.

Intentional Realism and Causation by Content

Intentional realism treats beliefs and desires as the sorts of things that can cause behavior. But this is a special type of causation. There is a fundamental difference between my leg moving because I am trying to achieve something (perhaps the journey of a thousand miles that starts with a single step) and my leg moving because a doctor has hit my knee with his hammer. In the first case, what causes my movement is what the desire is a desire for, namely, the beginning of the journey of a thousand miles. This is what philosophers call the *content* of the desire. There is nothing corresponding to a desire with content (or any other state with content) when a doctor hits my knee with a hammer. The movement that I make is simply a

response to a physical stimulus. It is not a response to something that I want to achieve.

This phenomenon, often called *causation by content* , is something that any version of intentional realism has to explain. Yet, on the face of it, causation by content is deeply mysterious. It depends on representations (stored information about the environment). In one sense, representations are simply objects like any other — they might be patterns of sound waves, populations of neurons, or pieces of paper. Thought of in this way, it is no more difficult to understand how representations can cause behavior than it is to understand how the doctor ' s hammer can make my leg move.

But the representations that we are interested in (such as beliefs and desires) are also things that bear a special *semantic* relation to the world — they have meanings. And the effects that representations have in the world are a function of what they mean. But how can representations have causal effects within the world in virtue of their semantic properties?

Fodor, along with almost all cognitive scientists and the vast majority of philosophers, holds that brains and the representations that they contain are physical entities. This means that brains can only be sensitive to the physical properties of mental representations. My utterance of the word *cat* is ultimately no more than a particular pattern of sound waves. These sound waves have certain physical properties (amplitude, frequency, wavelength, and so on) that can have certain effects on the brain. But the fact that those sound waves represent cats for English speakers is not a physical property (or so the argument goes).

Let us call the physical properties that can be manipulated within brains *formal properties.* They have to do with the physical *form* (i.e., the shape) of the representation. And let ' s use *semantics* for the properties that enable

representations to represent — just as semantics is the branch of linguistics that deals with the meanings of words (how words represent).

This gives us another way of putting the problem. How can the brain be an information-processing machine if it is blind to the semantic properties of representations? How can the brain be an information-processing machine if all it can process are the formal properties of representations?

Exercise 4.4 Explain the contrast between formal and semantic properties in your own words.

Think about this in terms of computers. Computers essentially manipulate strings of symbols. A computer programmed in binary, for example, manipulates strings of 1s and 0s. This string of 1s and 0s might represent a natural number, in the way that in binary, 10 represents the number 2, and 11 represents the number 3. Or it might represent something completely different. It might represent whether or not the individual members of a long series of pixels are on or off, for example.

In fact, with suitable coding, a string of 1s and 0s can represent just about anything. As far as the computer is concerned, however, what the string of 1s and 0s represents is completely irrelevant. The semantic properties of the string are irrelevant. The computer simply manipulates the formal properties of the string of 1s and 0s. In fact, it would be more accurate to say that the computer operates on *numerals* rather than *numbers.* Numerals are just symbols with particular shapes. Numbers are what those numerals represent.

But here ' s where the computer program comes in. The computer is programmed to manipulate strings of 1s and 0s in certain ways that yield the right result, even though the computer has no idea what that right result is. Take an adding machine, for example. Suppose it is given two strings of 0s and 1s and, in response, outputs a third string of 1s and 0s. If the first two strings represent the numbers 5 and 7, respectively, then (if the machine is well designed) the third string will be a binary representation of the number 12.

But even though all the computer is doing is mechanically manipulating 1s and 0s (numerals, not numbers), operating on their formal properties, it nonetheless comes up with the right answer. So, although the computer itself is not concerned with what " 12 " means, the computer program must respect the rules of addition in order for the computational result —" 12 "— to have meaning.

In essence, what computer programmers do when they are programming an adding machine is writing code so that purely mechanical manipulations of numerals will correctly track the arithmetical relations between the numbers that the numerals represent. So, the adding machine must manipulate the numerals " 7 " and " 5 " in such a way that taking them as inputs to the machine yields the numeral " 12 " because it is an arithmetical fact that 7 + 5 = 12 (which is a statement about numbers, not numerals).

Fodor thinks that this is a great model for the human brain. Brains are physical systems that can be sensitive only to the formal properties of mental representations. But nonetheless, as information-processing machines, they (like computers) have to respect the semantic properties of mental representations. The LoT is what makes this possible.

Exercise 4.5 Explain the analogy between brains and computers in your own words.

The Language of Thought and the Relation between Syntax and Semantics

Here are the three main claims of Fodor ′s LoT hypothesis:

1. Causation through content takes place through causal interactions between physical states.

2. These physical states have the structure of sentences, and their sentence-like structure determines how they are made up and how they interact with each other.

3. Causal transitions between sentences in the LoT respect the semantic relations between the meanings of those sentences.

According to Fodor, we think in sentences, but not in sentences of a natural language such as English. The LoT is much closer to a logical language, such as the propositional calculus (which we looked at briefly earlier in this chapter — see Box 4.1). It is supposed to be free of the nuances, ambiguities, and multiple layers of meaning that we find in English and other natural languages.

Both the LoT and logical languages incorporate a clear separation between *syntax* and *semantics.* Syntax has to do with symbols and the rules for combining them. It is the logical equivalent of grammar. Semantics, on the other hand, has to do with what the symbols actually mean and, relatedly, to what makes sentences true (or false).

To illustrate this general distinction, we can use the *predicate calculus.* This is a logical language more powerful and sophisticated than the propositional calculus we looked at in Box 4.1 . Unlike the propositional calculus (which only allows us to talk about complete sentences or propositions), the predicate calculus allows us to talk directly about individuals and their properties. So, for example, the predicate calculus allows us to formalize inferences such as the following:

Hubert is laughing.

Therefore, someone is laughing.

Or:

Everyone is laughing.

Therefore, Hubert is laughing.

In order to represent these inferences, the predicate calculus has special symbols. These include individual constants that name particular objects; predicate letters that name properties; and *quantifiers* , which correspond to the English words *some* and *all* . The symbols are typically identifiable by simple typographical features (such as uppercase for predicate letters and lowercase for individual constants), and they can be combined to make complex symbols according to certain rules.

From a syntactic point of view, the predicate calculus is simply a set of symbols of various types together with rules for manipulating those symbols according to their types. These rules identify the symbols only in terms of their typographical features. An example would be the rule that the space after an uppercase letter (e.g., the space in " F –") can be filled only with a lowercase letter (e.g., " *a* "). This rule is a way of capturing, at the syntactic

level, the intuitive thought that properties apply primarily to things — because uppercase letters (such as "F–") can only be names of properties, whereas lowercase letters (such as "*a*") can only be names of objects. The rule achieves this, however, without explicitly stating anything about objects and properties. It is purely syntactic.

In contrast, when we think about the semantics of a formal language, we assign objects to the individual constants and properties to the predicates. We identify the particular object that each individual constant names, for example. To provide a semantics for a language is to give an interpretation to the symbols it contains — to turn it from a collection of meaningless symbols into a representational system.

Exercise 4.6 Explain the distinction between syntax and semantics in your own words.

Fodor thinks that sentences in the LoT behave the same way. They can be viewed purely syntactically — as physical symbol structures composed of basic symbols arranged according to certain rules of composition. Or they can be viewed semantically in terms of how they represent the world (in which case they are being viewed as the vehicles of propositional attitudes).

Recall that causal transitions between sentences in the LoT are essentially syntactic, that is, sensitive only to the formal properties of the relevant symbols, regardless of the symbols ' meanings. Then we need to ask:

> Why do the syntactic relations between sentences in the LoT map onto the semantic relations holding between the contents of those sentences?

If the LoT really is a formal system, then this question has a straightforward answer. Syntactic transitions between sentences in the LoT track semantic transitions between the contents of those sentences for precisely the same reason that syntax tracks semantics in any properly designed formal system.

Fodor can (and does) appeal to well-known results in meta-logic (the study of the expressive capacities and formal structure of logical systems). These results establish a significant degree of correspondence between syntax and semantics. So, for example, it is known that the first-order predicate calculus is sound and complete. That is to say, in every well-formed proof in the first-order predicate calculus, the conclusion really is a logical consequence of the premises (*soundness*), and conversely, for every argument in which the conclusion follows logically from the premises and both conclusion and premises are formulable in the first-order predicate calculus, there is a well-formed proof (*completeness*).

The combination of soundness and completeness has the following important consequences. If a series of legitimate syntactic transformations leads from formula A to a second formula B, then one can be sure that A cannot be true without B being true — and conversely, if A entails B in a semantic sense, then one can be sure that there will be a series of formally definable inferential transitions leading from A to B.

Here ' s an example. Suppose that we have two complex symbols, " F *a* " and " G *a* . " Each of these symbols is a sentence in the LoT with a particular syntactic shape. We know that " F –" and " G –" are symbols for *predicates.* Let us say that " F –" means "– is tall " and " G –" means "– has red hair. " We also know that " *a* " is a *name* symbol. Let us say that " *a* " names Georgina. The meaning of " F *a* " is that Georgina is tall, whereas the meaning of " G *a* " is that Georgina has red hair.

Table 4.1 shows how a very simple piece of thinking might be analyzed by the LoT hypothesis. It also shows how two physical symbols, " F *a* " and " G *a* , " are transformed in two inferential steps into the more complex physical symbol "∃ *x* (F *x* & G *x*). " Here "∃ *x* –" is the symbol for " there is at least one *x* such that – , " so this sentence means " There is at least one thing that is both F and G. "

Table 4.1 Syntax and semantics in the predicate calculus

SYMBOLS	TRANSFORMATION RULE	MEANING
1. F*a*		1. Georgina is tall.
2. G*a*		2. Georgina has red hair.
3. (F*a* & G*a*)	If complex symbols "S" and "T" appear on earlier lines, then it is legitimate to write "(S & T)"	3. Georgina is tall and has red hair.
4. ∃*x* (F*x* & G*x*)	If there is a complex symbol containing a name symbol on an earlier line, then it is legitimate to replace the name symbol with "*x*" and write "∃*x* –" in front of the complex symbol (NOTE: "∃*x*–" is the symbol for "there is at least one *x* such that –.")	4. At least one person is tall and has red hair.

The rules that achieve this transformation are purely syntactic. They are simply rules for manipulating symbol structures. But when we look at the relation between the meanings of " F *a* " and " G *a* , " on the one hand, and the meaning of "∃ *x* (F *x* & G *x*), " on the other, we see that those purely syntactic transformations preserve the logical relations between the propositions that the symbols stand for. If it is true that Georgina is tall and that Georgina has red hair, then it is certainly true that at least one person is tall and has red hair.

In sum, beliefs and desires are realized by language-like physical structures (sentences in the LoT), and practical reasoning and other forms of thinking are ultimately just causal interactions between those structures. These causal interactions are sensitive only to the formal, syntactic properties of the physical structures. Yet, because the LoT is a formal language with analogs of the formal properties of soundness and

completeness, these purely syntactic transitions respect the semantic relations between the contents of the relevant beliefs and desires. This is how (Fodor claims) causation by content can take place in a purely physical system such as the human brain. And so, he argues, the commonsense psychological explanation is vindicated by thinking of the mind as a computer processing sentences in the LoT.

Fodor ' s argument for the hypothesis is very complicated. To make it easier to keep track of the different steps, I have represented them diagrammatically in Figure 4.2 .

Exercise 4.7 Use the flowchart in Figure 4.2 to explain Fodor ' s argument in your own words.

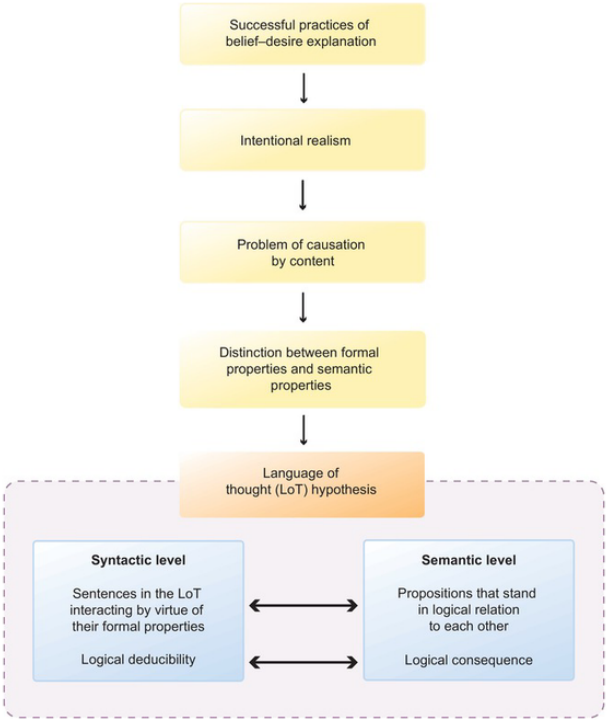


Figure 4.2 The structure of Fodor ' s argument for the language of thought hypothesis.

4.3 The Russian Room Argument and the Turing Test

We turn now to a fundamental objection to the physical symbol system hypothesis. This objection comes from philosopher John Searle, who is convinced that no machine built according to the physical symbol system hypothesis could possibly be capable of intelligent behavior. What he tries to do is describe a situation in which symbols are correctly manipulated, but where he argues that there is neither understanding nor intelligence.

Searle asks us to imagine a machine that he calls a Russian room. (Searle originally used the example of Chinese for dramatic effect because it is spoken and understood by very few people in the United States and is notoriously difficult to learn as a second language. I have decided to switch the example to Russian, for variety. Russian speakers should feel free to substitute their own example — an English room, for example.) Inside the room, a person receives pieces of paper through one window and passes out pieces of paper through another window. The pieces of paper have symbols in Russian written on them. The Russian room, in essence, is an input – output system, with symbols as inputs and outputs. The way it works is determined by a huge instruction manual that tells the person in the room which pieces of paper to pass out depending on which pieces of paper the person receives.

The instruction manual is essentially just a way of pairing input symbols with output symbols. It is not written in Russian and can be understood and followed by someone who knows no Russian. All that the person needs to be able to do is to identify Russian symbols in some sort of syntactic way — according to their shape, for example. This is enough for

them to be able to find the right output for each input — where the right output is taken to be the output dictated by the instruction manual.

Now, Searle continues, imagine first that the instruction manual has been written in such a way that the inputs are all questions in Russian and the outputs are all appropriate answers to those questions. For all intents and purposes, therefore, the Russian room is answering questions in Russian. Now imagine that the person in the room does not, in fact, know any Russian. All he is doing is following the instructions in the instruction manual (which is written in English) to match an outgoing piece of paper (the answer) to an incoming piece of paper (the question).

The situation is illustrated in Figure 4.3 . What the Russian room shows, according to Searle, is that it is perfectly possible for there to be syntactic symbol manipulation without any form of intelligence or understanding. The Russian room as a whole does not understand Russian because the man inside the room does not understand Russian.



Figure 4.3 Inside and outside the Russian room.

The Russian room seems to be set up in accordance with the physical symbol system hypothesis. The person in the Russian room is manipulating symbols according to their formal/syntactic properties, and it has been set up to produce the right output for every input. In the terms we used in the last section, the syntactic manipulation of the symbols preserves their semantic properties. The semantic properties of the input symbols are their meanings — that is, certain questions. The semantic properties of the output symbols are answers to those questions. So, as long as the person in the Russian room follows the instructions correctly, the semantic relations between input and output will be preserved.

And yet, Searle argues, the Russian room does not understand Russian. How can it understand Russian, given that the person in the room does not understand Russian?

To someone outside the room, it might look as if there is intelligent behavior going on. The machine does, after all, respond to the questions it is asked with answers that make sense. But this is just an illusion of intelligence. The Russian room cannot be behaving intelligently if it does not understand Russian. And so, it is a counterexample to the physical symbol system hypothesis — or so Searle argues.

In fact, Searle also thinks that the Russian room argument reveals a fundamental problem with the so-called Turing Test, proposed by Alan Turing in 1950 as a criterion for whether a machine was displaying real intelligence. Turing ′s basic idea is that if an observer is communicating with a machine and cannot tell the difference between it and a human being, then that would show that the computer was genuinely intelligent. You might imagine, for example, that a judge is communicating simultaneously with the machine and with a human being via a computer screen and keyboards but cannot tell which is which, however complicated and lengthy the interaction. For Turing, then, a machine that responds in exactly the way that a human being responds thereby counts as intelligent.

Exercise 4.8 Explain in your own words why the Russian room argument is an objection to taking the Turing Test to reveal genuine intelligence.

Responding to the Russian Room Argument

Many people have suggested that there is a crucial equivocation in the argument. The physical symbol system hypothesis talks about what makes

cognitive *systems* intelligent. But the crucial step in Searle ′ s argument is not a claim about the system as a whole. It is a claim about part of the system, namely, the person inside the room who is reading and applying the instruction manual.

According to the *systems reply* , the Russian room argument is simply based on a mistake about where the intelligence is supposed to be located. Supporters of the systems reply hold that the Russian room as a whole understands Russian and is displaying intelligent behavior, even though the person inside the room does not understand Russian.

Searle has a clever response. Instead of imagining yourself in the Russian room, imagine the Russian room inside you! If you memorize the instruction manual, then, Searle says, you have effectively internalized the Russian room. Of course, it ′ s hard to imagine that anyone could have a good enough memory to do this, but there are no reasons to think that it is impossible in principle. But, Searle argues, internalizing the Russian room in this way is not enough to turn you from someone who does not understand Russian into someone who does. After all, what you ′ ve memorized is not Russian but just a complex set of rules for mapping some symbols you don ′ t understand onto other symbols you don ′ t understand.

Exercise 4.9 How might a defender of the systems reply respond to Searle ′ s response?

Another possible response to the Russian room argument is the *robot reply.* The robot reply grants Searle ′ s basic claim that the Russian room does not understand Russian but offers a different explanation. The problem

with the Russian room has nothing to do with the problem of getting from syntax to semantics. The problem, rather, is that it is embodied agents who understand Russian, not disembodied cognitive systems into which pieces of paper enter and from which other pieces of paper come out.

The ability to understand Russian involves, at a minimum, being able to carry out instructions given in Russian, coordinate with other Russian speakers, read Russian characters, and carry on a conversation. In order to build a machine that could do all this, we would need to embed the Russian room in a robot, providing it with some analog of sensory organs, vocal apparatus, and limbs. If the Russian room had all this and could behave in the way that a Russian speaker behaves, then, a supporter of the robot reply would say, there is no reason to deny that the system understands Russian and is behaving intelligently.

Again, Searle is unconvinced. For him, the gulf between syntax and semantics can ′ t simply be overcome by turning the Russian room into a Russian robot. An embodied Russian room might indeed stop when it " sees " the Russian word for " stop. " But this would simply be something it has learned to do. It no more understands what the character means than a laboratory pigeon that has been trained not to peck at a piece of card with the same character on it. Interacting with the environment is not the same as understanding it. Even if the Russian robot does and says all the right things, this does not show that it understands Russian. The basic problem still remains, as far as Searle is concerned: simply manipulating symbols cannot create meaning, and unless the symbols are meaningful to the Russian room, there is no relation between what it does and what a " real " Russian speaker might do.

Exercise 4.10 Explain the robot reply in your own words and assess Searle ′ s response to it.

The Russian room argument raises fundamental questions about the nature of intelligence and, relatedly, of what counts as genuine thinking. Searle often presents it as an objection to what is often called the project of *strong AI* — this is the project of building machines that are genuinely intelligent (as opposed to the project of *weak AI* , which simply aims to build machines that do the things that human beings can do). And this is what many supporters (and critics) of the Russian room argument have focused on.

Moreover, if the argument is sound, it strikes at the very possibility of the physical symbol system hypothesis. Searle tries to show that there must be more to genuine thinking than simply manipulating symbols according to rules, whereas the physical symbol system hypothesis says that this is all that thinking is.

Summary

This chapter has looked at Newell and Simon ′ s claim (the physical symbol system hypothesis) that thinking consists of manipulating symbol structures according to rules. After introducing the physical symbol system hypothesis, we considered Jerry Fodor ′ s suggestion that these symbol structures are sentences in an internal language of thought. The chapter ended with the

Russian room argument, an objection to the basic idea of the physical symbol system hypothesis.

Checklists

The physical symbol system hypothesis states that a physical symbol system has the necessary and sufficient means for general intelligent action. In more detail:

1. The symbols are physical patterns.

2. Physical symbols can be combined to form complex symbol structures.

3. Physical symbol systems contain processes for manipulating complex symbol structures.

4. The processes for manipulating complex symbol structures can be represented by symbols and structures within the system.

5. Problems are solved by generating and modifying symbol structures until a solution structure is reached.

Fodor ' s language of thought (LoT) hypothesis is one way of turning the physical symbol system hypothesis into a concrete proposal about how the mind works.

1. The LoT hypothesis is grounded in intentional realism. Psychological states such as belief and desire are real physical entities. These entities are sentences in the LoT.

2. The hypothesis offers a way of explaining causation by content (i.e., how physical representations can have causal effects in the world as a function of how they represent the world).

3. Fodor suggests that we understand the relation between sentences in the LoT and their contents based on the model of the relation between syntax and semantics in a formal system.

4. The syntax of the LoT tracks its semantics because it is a formal language with analogs of the formal properties of soundness and completeness.

The Russian room argument is a thought experiment directed against the idea that the rule-governed manipulation of symbols is sufficient to produce intelligent behavior.

1. The person in the Russian room is manipulating symbols according to their formal/syntactic properties, without any understanding of Russian.

2. According to the systems reply, the real question is whether the system as a whole understands Russian, not whether the person in the room understands Russian.

3. According to the robot reply, the Russian room does not understand Russian. But this is not because of any uncrossable gap between syntax and semantics. Rather, it is because the Russian room has no opportunity to interact with the environment and other people.

Further Reading

The paper by Newell and Simon discussed in Section 4.1 is reprinted in a number of places, including Boden 1990b and Bermúdez 2006 . A good introduction to the general ideas behind the physical symbol system hypothesis in the context of AI is Haugeland 1985 , particularly Chapter 2, and Haugeland 1997 , Chapter 4. See also Chapters 1 – 3 of Johnson-Laird 1988 , Chapters 4 and 5 of Copeland 1993 , Chapter 2 of Dawson 1998 , and the *Encyclopedia of Cognitive Science* entry on symbol systems (Nadel 2005). Russell and Norvig 2009 is the third edition of a popular AI textbook. Also see Poole and Mackworth 2010 , Warwick 2012 , and Proudfoot and Copeland ' s chapter on AI in *The Oxford Handbook of Philosophy of Cognitive Science* (Margolis et al. 2012).

Fodor 1975 and 1987 are classic expositions on the LoT approach from a philosophical perspective. See also Fodor 2008 . For a psychologist ' s perspective, see Pylyshyn ' s book *Computation and Cognition* (1984) and his earlier target article in *Behavioral and Brain Sciences* (1980). More recent philosophical discussions of the LoT can be found in Schneider 2011 and Schneider and Katz 2012 . The *Encyclopedia of Cognitive Science* (Shanahan 2003) has an entry on the topic, as does the *Stanford Encyclopedia of Philosophy* (https://plato.stanford.edu/). For a general philosophical discussion of the computational picture of the mind, Crane 2003 and Sterelny 1990 are recommended. Block 1995a explores the metaphor of the mind as the software of the brain. Fodor ' s argument for the LoT hypothesis is closely tied to important research in mathematical logic and the theory of computation. Rogers 1971 is an accessible overview. For

general introductions to philosophical debates about mental causation and the more general mind－body problem, see Heil 2004 and Searle 2004 .

Searle originally presented his argument as a target article in the journal *Behavioral and Brain Sciences* , with extensive commentary from many cognitive scientists (Searle 1980). Margaret Boden ′ s article " Escaping from the Chinese Room " (1990a), reprinted in Heil 2004 , is a good place to start in thinking about the argument. In January 1990, the periodical *Scientific American* devoted a special issue to the tenth anniversary of Searle ′ s argument. Churchland and Churchland 1990 was one of the contributions, arguing that although the argument is effective against classical AI, it leaves artificial neural networks untouched. Preston and Bishop 2002 is a collection of articles dedicated to the Russian room argument, covering the principal lines of response. The entry on the Chinese room argument in the online *Stanford Encyclopedia of Philosophy* (https://plato.stanford.edu/) is comprehensive and has a very full bibliography. A related problem, the symbol-grounding problem, is introduced and discussed in Harnad 1990 (see the online resources). Taddeo and Floridi 2005 critically reviews a number of proposed solutions.

Chapter 5

Neural Networks and Distributed Information Processing

Overview

5.1 Neurally Inspired Models of Information Processing

Neurons and Network Units

5.2 Single-Layer Networks and Boolean Functions

Learning in Single-Layer Networks: The Perceptron-Convergence Rule

Linear Separability and the Limits of Perceptron Convergence

5.3 Training Multilayer Networks

The Backpropagation Algorithm

How Biologically Plausible Are Neural Networks?

5.4 Information Processing in Neural Networks: Key Features

Distributed Representations

No Clear Distinction between Information Storage and Information Processing

The Ability to Learn from " Experience "

Overview

This chapter looks at neural network models of information processing. The chapter begins with Section 5.1 , which reviews some of the motivations for neurally inspired models of information processing and looks at how the individual units in neural networks compare to biological neurons.

The simplest artificial neural networks are single-layer networks. These are explored in Section 5.2 . We will see that any digital computer can be simulated by a suitably chained-together set of single-layer networks. However, these networks are limited in what they can *learn* .

Overcoming those limitations requires moving from single-layer networks to multilayer networks, which are capable of learning through strategies that reduce error. In Section 5.3 , we look at the backpropagation algorithm used to train multilayer networks. Finally, Section 5.4 summarizes the key features of information processing in multilayer artificial neural networks, explaining key differences between neural networks and physical symbol systems.

5.1 Neurally Inspired Models of Information Processing

Detailed knowledge of how the brain works has increased dramatically in recent years (see Chapter 3). Neuroimaging techniques, such as functional magnetic resonance imaging (fMRI) and positron emission tomography (PET), have allowed neuroscientists to begin establishing large-scale correlations between types of cognitive functioning and specific brain areas. Combining this with the information available from studies of patients with

brain damage allows cognitive scientists to build up a functional map of the brain.

Other techniques have made it possible to study brain activity (in nonhuman animals, from monkeys to sea slugs) at the level of the single neuron. Microelectrodes can be used to record electrical activity both inside a single neuron and in the vicinity of that neuron. Recording from inside neurons allows a picture to be built up of the different types of input to the neuron, both excitatory and inhibitory, and the mechanisms that modulate output signals.

But none of these techniques offers direct insight into *how* information is processed in the brain. PET and fMRI tell us which brain areas are involved in particular cognitive tasks but not how those cognitive tasks are actually carried out. Nor will this information come from single-neuron recordings. We may well find out from single-neuron recordings in monkeys that particular types of neurons in particular areas of the brain respond very selectively to a narrow range of visual stimuli, but as yet, we have no idea how to scale this up into an account of how vision works.

The brain is an extraordinarily complicated set of interlocking and interconnected circuits. Its most fundamental feature is its *connectivity* , and the crucial question in understanding the brain is how distributed patterns of activation across populations of neurons can give rise to perception, memory, sensorimotor control, and high-level cognition. But we have (as yet) limited tools for directly studying how populations of neurons work.

Because we do not have the equipment and resources to study populations of neurons directly, many researchers have developed techniques for studying populations of neurons indirectly. One strategy is

constructing models that approximate populations of neurons. These are called *neural network models* or *artificial neural networks* .

There are many different types of neural network models and many different ways of using them. The focus in *computational neuroscience* is on modeling biological neurons and populations of neurons. Computational neuroscientists start from what is known about the biology of the brain and then construct models by abstracting away from some biological details while preserving others. *Connectionist modelers* often pay less attention to the constraints of biology. They tend to start with generic models. Their aim is to show how those models can be modified and adapted to simulate and reproduce well-documented psychological phenomena, such as the patterns of development that children go through when they acquire language or the way in which cognitive processes break down in patients with brain damage.

For our purposes here, the differences between computational neuroscientists and connectionist modelers are less important than what they have in common. Neural network models are distinctive in how they store information, how they retrieve it, and how they process it. And even those models that are not biologically driven remain neurally inspired.

Neurons and Network Units

Neural networks are made up of individual units loosely based on biological neurons. The many different types of neurons in the nervous system all share a common basic structure. Each neuron is a cell and so has a cell body (a *soma*) containing a nucleus. There are many root-like extensions from the cell body. These are called *neurites.* There are two different types of neurites. Each neuron has many dendrites and a single axon. The dendrites

are thinner than the axon and form what looks like a little bush (as illustrated in Figure 5.1). The axon itself eventually splits into a number of branches, each terminating in a little *endbulb* that comes close to the dendrites of another neuron .

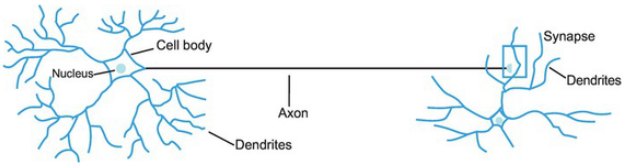


Figure 5.1 Schematic illustration of a typical neuron.

Neurons receive signals from other neurons. A typical neuron might receive inputs from 10,000 neurons, but the number is as great as 50,000 for some neurons in the brain area called the *hippocampus* . These signals are received through the dendrites, which can be thought of as the receiving end of the neuron. A sending neuron transmits a signal along its axon to a *synapse* , which is the site where the end of an axon branch comes close to a dendrite or the cell body of another neuron. When the signal from the sending (or *presynaptic*) neuron reaches the synapse, it generates an electrical signal in the dendrites of the receiving (or *postsynaptic*) neuron.

The basic activity of a neuron is to fire an electrical impulse along its axon. The single most important fact about the firing of neurons is that it depends on activity at the synapses. Some of the signals reaching the neuron ′ s dendrites promote firing, and others inhibit it. These are called *excitatory* and *inhibitory* synapses, respectively. If the sum of excitatory and inhibitory synapses exceeds the *threshold of the neuron* , then the neuron will fire.

Neural networks are built up of interconnected populations of units that share some of the generic characteristics of biological neurons. Figure 5.2 illustrates a typical network unit. There are n inputs, corresponding to

synaptic connections to presynaptic neurons. Signals from the presynaptic neurons might be excitatory or inhibitory. This is captured in the model by assigning a numerical weight W_i to each input I_i. Typically, the weight will be a real number between 1 and − 1. A positive weight corresponds to an excitatory synapse, and a negative weight corresponds to an inhibitory synapse.

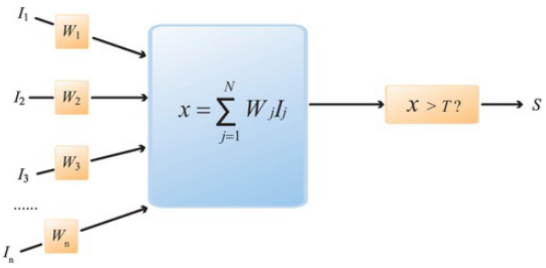


Figure 5.2 An artificial neuron.

Multiplying each input by its weight indicates the strength of the signal at each synapse. Adding all these individual *activation levels* together gives the total input to the unit, corresponding to the total signal reaching the nucleus of the neuron. This is represented using standard mathematical format in Figure 5.2 . (A reminder: Σ is the symbol for summation [repeated addition]. The N above the summation sign indicates that there are N things to add together. Each of the things added together is the product of I_j and W_j for some value of j between 1 and N .) If the total input exceeds the threshold (T), then the neuron " fires " and transmits an output signal .

The one thing that remains to be specified is the strength of the output signal. We know that the unit will transmit a signal if the total input exceeds its designated threshold, but we do not yet know what that signal is. For this, we need to specify an *activation function* — a function that assigns an output signal on the basis of the total input. Neural network designers standardly

choose from several different types of activation functions. Some of these are illustrated in Figure 5.3 .

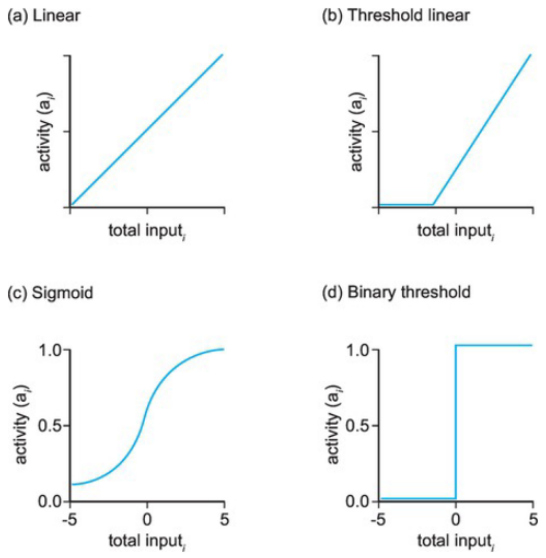


Figure 5.3 Four different activation functions. Each one fixes a neuron ′s activation level as a function of the total input to the neuron.

(Adapted from McLeod et al. 1998 .)

The simplest is a linear function in which the output signal increases in direct proportion to the total input. (Linear functions are so called because they take a straight line when drawn on a graph.) The threshold linear function is a slight modification of this. This function yields no output signal until the total input reaches the threshold — and then the strength of the output signal increases proportionately to the total input. There is also a binary threshold function, which effectively operates like an on/off switch. It either yields zero output (when the input signal is below threshold) or maximum output (when the input signal is at or above threshold).

The threshold functions are intended to reflect a very basic property of biological neurons, which is that they only fire when their total input is

suitably strong. The binary threshold-activation function models neurons that either fire or don ′ t fire, whereas the threshold linear function models neurons whose firing rate increases in proportion to the total input once the threshold has been reached.

The sigmoid function is a very commonly used *nonlinear* activation function. This reflects some of the properties of real neurons in that it effectively has a threshold below which total input has little effect and a ceiling above which the output remains more or less constant despite increases in total input. The ceiling corresponds to the maximum firing rate of the neuron. Between the threshold and the ceiling, the strength of the output signal is roughly proportionate to the total input and so looks linear. But the function as a whole is nonlinear and drawn with a curve .

The next step is to see how these individual units can be used to process information. We will start out by looking at the simplest neural networks. These are *single-layer networks.*

5.2 Single-Layer Networks and Boolean Functions

The first neural networks were studied in the 1940s and 1950s, pioneered by the neuroscientist Warren McCulloch and the logician Walter Pitts. They were known as *single-layer networks* . To see what single-layer networks can do (and what they can ′ t do), we need to start with a quick refresher on *mapping functions.*

The basic idea of a function should be familiar, even if the terminology may not be. Addition is a function. Given two numbers as *inputs* , the

addition function yields a third number as *output.* The output is the sum of the two inputs. Multiplication is also a function. Here, the third number is the product of the two inputs.

Some helpful terminology. Suppose that we have a set of items. We can call that a *domain* . Let there be another set of items, which we can call the *range.* A mapping function maps each item from the domain onto exactly one item from the range. The defining feature of a function is that no item in the domain gets mapped to more than one item in the range. Functions are *single-valued.* The operation of taking square roots, for example, is not a function (at least when negative numbers are included) because every positive number has two square roots. Figure 5.4 gives an example of a mapping function.

Exercise 5.1 Give another example of an arithmetical operation that counts as a function, as well as another example of an operation that is not a function.

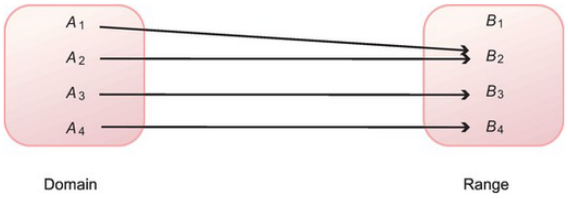


Figure 5.4 Illustration of a mapping function. A mapping function maps each item in its domain to exactly one item in its range.

Let ′ s turn now to mapping functions of a very special kind. These are functions with a range consisting of two items, one corresponding to TRUE and the other corresponding to FALSE. These functions are ways of

classifying objects in the domain. If the domain of the function contains all the natural numbers and the range of the function contains two items corresponding to TRUE and FALSE, then we can identify any subset we please of the natural numbers by mapping the members of that subset onto TRUE and all the others onto FALSE. If the subset that the function maps onto TRUE contains all and only the even numbers, for example, then we have a way of picking out the set of the even numbers.

Now, we have all the machinery we need to introduce the so-called *binary Boolean functions.* These functions all have the same range as our even-number function, namely, the set consisting of the two truth values TRUE and FALSE. Instead of having numbers in the domain, however, the domain of these functions is made up of pairs of truth values.

There are four different possible pairs of truth values. These pairs form the domain of the binary Boolean functions. The range, as with all Boolean functions, is given by the set {TRUE, FALSE}, illustrated as follows:

DOMAIN	RANGE
FALSE, FALSE	
FALSE, TRUE	FALSE
TRUE, FALSE	TRUE
TRUE, TRUE	

Each binary Boolean function assigns either TRUE or FALSE to each pair of truth values. You can think of a binary Boolean function as a way of showing how to fix the truth value of a complex sentence built up from two simpler sentences on the basis of the truth values of those simpler sentences.

Some Boolean functions should be very familiar. There is a binary Boolean function standardly known as AND, for example. AND maps the pair {TRUE, TRUE} to TRUE and maps all other pairs of truth values to

FALSE. To put it another way, if you are given a sentence A and a sentence B, then the only circumstance in which it is true to claim A AND B is the circumstance in which both A and B have the value TRUE.

Similarly, OR is the name of the Boolean function that maps the pair {FALSE, FALSE} to FALSE and the other three pairs to TRUE. Alternatively, if you are given sentences A and B, then the only circumstance in which it is false to claim A OR B is the circumstance in which both A and B have the value FALSE.

It is important that the OR function assigns TRUE to the pair {TRUE, TRUE}, so A OR B is true in the case where both A and B are true. As we shall see, there is a Boolean function that behaves just like OR, except that it assigns FALSE to {TRUE, TRUE}. This is the so-called *XOR function* (an abbreviation of exclusive-OR). XOR *cannot* be represented by a single-layer network. We will look at this in more detail later in this section.

We can represent these functions using *truth tables* . The truth table for AND tells us how the truth value of A AND B varies according to the truth value of A and B, respectively (or, as a logician would say, as a *function* of the truth values of A and B).

A	B	A AND B
FALSE	FALSE	FALSE
FALSE	TRUE	FALSE
TRUE	FALSE	FALSE
TRUE	TRUE	TRUE

Exercise 5.2 Give a truth table for the Boolean function OR.

Now, back to neural networks. McCulloch and Pitts observed that very simple neural networks can represent some binary Boolean functions. The first step is to represent Boolean functions using numbers (because we need numbers as inputs and outputs for the arithmetic of the activation function to work). This is easy. We can represent TRUE by the number 1 and FALSE by 0, as is standard in logic and computer science.

It is easy to see how we design our network unit to take only 0 and 1 as input. But how do we design it to produce only 0 and 1 as output?

This is where the activation functions come in. Binary threshold-activation functions output 0 until the threshold is reached, and then they output 1, irrespective of how the input increases. So, to represent a particular Boolean function, we need to set the weights and the threshold so that the network mimics the relevant truth table. A network that represents AND, for example, will have to output a 0 whenever the input is either (0, 0), (0, 1), or (1, 0). And it will have to output a 1 whenever the input is (1, 1).

The trick is setting the weights and threshold appropriately. Look at Figure 5.5 . If we set the weights at 1 for both inputs and the threshold at 2, then the unit will only fire when both inputs are 1. If both inputs are 1, then the total input is $(I_1 \times W_1) + (I_2 \times W_2) = (1 \times 1) + (1 \times 1) = 2$, which is the threshold. Because the network is using a binary threshold-activation function, in this case the output will be 1. If either input is a 0 (or both are), then the threshold will not be met, and so the output is 0. Because 1 represents TRUE and 0 represents FALSE, this network represents the AND function. It functions as what computer scientists call an *AND-gate* .

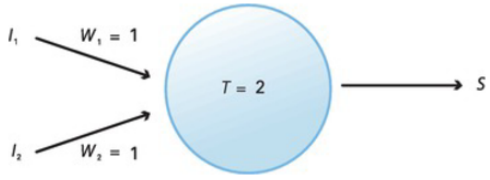


Figure 5.5 A single-layer network representing the Boolean function AND.

> **Exercise 5.3** Show how a network unit can represent OR and hence function as an OR-gate.

There are Boolean functions besides the binary ones. In fact, there are *n* -ary Boolean functions for every natural number *n* (including 0). But cognitive scientists are generally only interested in one nonbinary Boolean function. This is the unary function NOT, which is true if A is false and false if A is true. Again, this is easily represented by a single network unit, as illustrated in Figure 5.6 . The trick is to set the weights and threshold to get the desired result.

> **Exercise 5.4** Explain why the network unit in Figure 5.6 represents the unary Boolean function NOT.

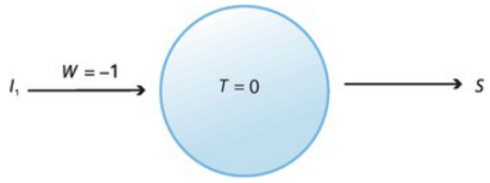


Figure 5.6 A single-layer network representing the Boolean function NOT.

So, single-layer networks can achieve a lot. As any computer scientist knows, modern digital computers are, in the last analysis, just incredibly complicated systems of AND-gates, OR-gates, and NOT-gates. So, by chaining together individual network units into a network, we can do anything that can be done by a digital computer. (This is why I earlier said that cognitive scientists are generally only interested in one nonbinary Boolean function. AND, NOT, OR, and a little ingenuity are enough to simulate any *n*- ary Boolean function, no matter how complicated.)

There is something missing, however. As we have seen, the key to getting single units to represent Boolean functions such as NOT and OR lies in setting the weights and the threshold. But this raises some fundamental questions: How do the weights get set? How does the threshold get set? Is there any room for learning?

Learning in Single-Layer Networks: The Perceptron- Convergence Rule

In 1949 Donald Hebb published *The Organization of Behavior* , in which he speculated about how learning might take place in the brain. His basic idea (the idea behind what we now call *Hebbian learning*) is that learning is, at bottom, an associative process. He famously wrote:

> When an axon of a cell A is near enough to excite cell B or repeatedly or persistently takes part in firing it, some growth or metabolic change takes place in both cells such that A ′ s efficiency, as one of the cells firing B, is increased.
>
> (Hebb 1949 : 11)

If A is a presynaptic neuron and B is a postsynaptic neuron, then every time that B fires after A fires increases the probability that B will fire after A fires (this is what Hebb means by an increase in A ′ s efficiency).

In slogan form, Hebbian learning is the principle that *neurons that fire together, wire together*. It has proved a very useful tool in modeling basic pattern recognition and pattern completion, as well as featuring in more complicated learning algorithms, such as the competitive learning algorithm discussed in Section 5.3 .

Hebb was speculating about real neurons, not artificial ones, but at around the same time, Frank Rosenblatt was studying learning in single-layer networks. In an influential article in 1958, Rosenblatt called these networks *perceptrons* and proposed a learning rule called the *perceptron-convergence rule* .

The perceptron-convergence rule has some similarities with Hebbian learning. Like Hebbian learning, it relies on the basic principle that changes in weight are determined solely by what happens locally — that is, by what happens at the input and what happens at the output. But unlike Hebbian learning, it is a supervised algorithm — it requires feedback about incorrect solutions to the problem the network is trying to solve.

The perceptron-convergence rule works by reducing error. We (as supervisors of the network) can measure the discrepancy between the output that the network actually produces and the output that it is supposed to produce. We can label that discrepancy δ (small delta). It will be a number — the number reached by subtracting the actual output from the correct output. So:

$$\delta = \text{INTENDED OUTPUT} - \text{ACTUAL OUTPUT}$$

Suppose, for example, that we are trying to produce a network that functions as an AND-gate. This means that when the inputs each have a value of 1, the desired output is 1 (because A AND B is true in the case where A is true and B is true). If the output that the network actually produces is 0, then $\delta = 1$. If, in contrast, the desired output is 0 and the actual output is 1, then $\delta = -1$.

It is standard when constructing neural networks to specify a learning rate. This is a constant number between 0 and 1 that determines how large the changes are on each trial. We can label the learning rate constant ε (epsilon). The perceptron-convergence rule is a very simple function of δ and ε .

If we use the symbol Δ (big delta) to indicate the adjustment that we will make after each application of the rule, then the perceptron-convergence rule can be written like this (remembering that T is the threshold, I_i is the i th input, and W_i is the weight attached to the i th input):

$$\Delta T = -\varepsilon \times \delta$$
$$\Delta W_i = \varepsilon \times \delta \times I_i$$

There are multiple adjustments here, one for the threshold and one for each of the weights. Let ′ s look at how the equations work to see how the network learns by making these adjustments.

Suppose δ is positive. This means that our network has undershot (because it means that the correct output is greater than the actual output). Because the actual output is weaker than required, we can make two sorts of changes in order to close the gap between the required output and the actual output. We can *decrease* the threshold, and we can *increase* the weights.

This is exactly what the perceptron-convergence rule tells us to do. We end up decreasing the threshold because when δ is positive, $-\varepsilon \times \delta$ is negative. And we end up increasing the weights because $\varepsilon \times \delta \times I_i$ comes out positive when δ is positive.

> **Exercise 5.5** How does the perceptron-convergence rule work when the network overshoots?

Consider the very simple single-layer network in Figure 5.7 . This network only takes one input, so we only have one weight to worry about. We can take the starting weight to be − 0.6 and the threshold to be 0.2. Let ′ s set our learning constant at 0.5 and use the perceptron-learning rule to train this network to function as a NOT-gate.



Figure 5.7 The starting configuration for a single-layer network being trained to function as a NOT-gate through the perceptron-convergence rule. It begins with a weight of − 0.6 and a threshold of 0.2.

Suppose that we input a 1 into this network (where, as before, 1 represents TRUE and 0 represents FALSE). The total input is $1 \times -0.6 = -0.6$. This is below the threshold of 0.2, so the output signal is 0. Because this is the desired output, we have $\delta = 0$, so no learning takes place (because $\Delta T = -\varepsilon \times \delta = -0.5 \times 0 = 0$, and ΔW also comes out as 0). But if we input a 0, then we get a total input of $0 \times -0.6 = 0$. Because

this is also below the threshold, the output signal is 0. But this is not the desired output, which is 1. So, we can calculate $\delta = 1 - 0 = 1$. This gives $\Delta T = -0.5 \times 1 = -0.5$ and $\Delta W = 0.5 \times 1 \times 0 = 0$. This changes the threshold (to – 0.3) and leaves the weight unchanged.

This single application of the perceptron-convergence rule is enough to turn the network into a NOT-gate. If we input a 1 into the network, then the total input is $1 \times -0.6 = -0.6$, which is below the threshold. So the output signal is 0, as required. And if we input a 0 into the network, then the total input is $0 \times -0.6 = 0$, which is above the threshold of – 0.3. So, the output signal is 1, as required. In both cases, we have $\delta = 0$, so no further learning takes place. The network has *converged* on a solution.

Linear Separability and the Limits of Perceptron Convergence

The perceptron-convergence rule is very powerful. In fact, it can be proved (although we won ' t do so here) that applying the rule is guaranteed to converge on a solution in every case in which a solution exists. But it has its limits, and there is a relatively simple way of classifying the functions that a network can and cannot learn to compute through perceptron convergence.

To see this, we can use a graph to represent the " space " of possible inputs into a function. Figure 5.8 shows how to do this for two functions. The function on the left is the AND function. On the graph, a black dot is used to mark the inputs for which the AND-gate outputs a 1, and a white dot marks the inputs that get a 0. There are four possible inputs and, as expected, only one black dot (corresponding to the case where both inputs have the value TRUE). The graph for AND shows that we can use a straight line to separate out the inputs that receive the value of 1 from the inputs that receive

the value of 0. Functions that have this property are said to be *linearly separable.*

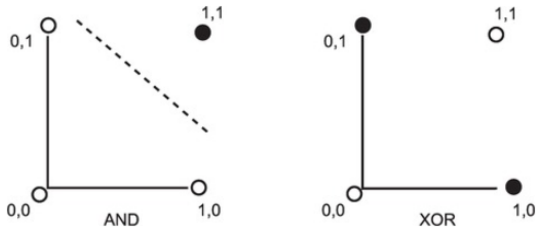


Figure 5.8 Graphical representations of the AND and XOR (exclusive-OR) functions, showing the linear separability of AND. Each of the four circles marked on the graph represents a possible combination of input truth values (as fixed by its coordinates). The circle is colored black only if the function outputs 1 at that point.

Exercise 5.6 Draw a graph to show that OR is linearly separable.

Clearly, though, the function on the right in Figure 5.8 is not linearly separable. This is the exclusive-OR function (standardly written as *XOR*). The OR function that we have been looking at up to now has the value TRUE even when both A and B have the value TRUE. This is not how the word *or* often works in English. If I am offered a choice between A or B, it often means that I have to choose one, but not both. This way of thinking about " or " is captured by the function XOR. A XOR B has the value TRUE only when exactly one of A and B has the value TRUE.

No straight line separates the black dots from the white dots in the graph of XOR. This means that XOR is not linearly separable. It turns out, moreover, that XOR cannot be represented by a single-layer network. This is easier to see if we represent XOR in a truth table. The table shows what the

output is for each of the four different possible pairs of inputs — as usual, 1 is the TRUE input and 0 is the FALSE input.

I_1	I_2	OUTPUT
0	0	0
0	1	1
1	0	1
1	1	0

Now, think about how we would need to set the weights and the threshold to get a single-layer network to generate the right outputs. We need the network to output a 1 when the first input is 0 and the second input is 1. This means that W_2 (the weight for the second input) must be such that $1 \times W_2$ is greater than the threshold — and likewise for the case where the first input is 1 and the second input is 0. In order to get this to come out right, we need W_1 to be such that $1 \times W_1$ is greater than the threshold. But now, with the weights set like that, it is inevitable that the network will output a 1 when both inputs are 1 — if each input is weighted so that it exceeds the threshold, then it is certain that adding them together will exceed the threshold. In symbols, if $W_1 > T$ and $W_2 > T$, then it is inevitable that $W_1 + W_2 > T$.

So, XOR fails to be linearly separable and is also not computable by a single-layer network. In fact, there is a general principle here. The class of Boolean functions that can be computed by a single-unit network is precisely the class of linearly separable functions. This was proved by Marvin Minsky and Seymour Papert in a very influential book entitled *Perceptrons* that was published in 1969.

But why does this matter? It is not too hard to construct an artificial network that will compute XOR. Figure 5.9 shows a network that will do the job. It is what is known as a *multilayer network.* Whereas the units in single-

layer networks receive inputs directly, multilayer networks contain *hidden units* that only receive inputs indirectly. The only inputs they can receive are outputs from other units.

Exercise 5.7 There are two binary Boolean functions that fail to be linearly separable. The second is the reverse of XOR, which assigns 1 where XOR assigns 0 and assigns 0 where XOR assigns 1. Construct a network that computes this function.

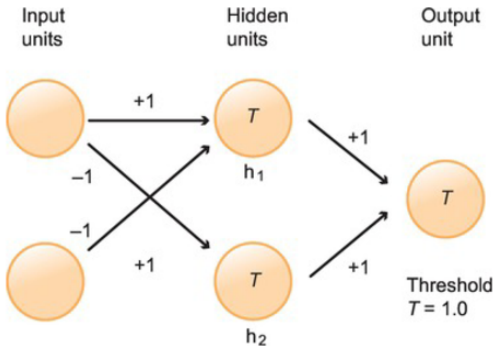


Figure 5.9 A multilayer network representing the XOR (exclusive-OR) function. Note that unlike the single-layer perceptrons that we have been considering up to now, this network has three layers. One of these layers is a hidden layer — it receives inputs only indirectly from other units.

(Adapted from McLeod et al. 1998 .)

The hidden units are what allows the network in Figure 5.9 to compute the XOR function. The reason a single-layer network cannot compute XOR is that it can only assign one weight to each input. This is why a network that outputs 1 when the first input is 1 and outputs 1 when the second input is 1 has to output 1 when both inputs are 1. This problem goes away when a network has hidden units. Each input now has its own unit, and each input

unit is connected to two different output units. This means that two different weights can now be assigned to each input.

Multilayered networks can compute any computable function — not just the linearly separable ones. But what stopped researchers in their tracks in 1969 was the fact that they had no idea how to train multilayered networks. The great breakthrough in neural network modeling came with the discovery of an algorithm for training multilayer networks.

> **Exercise 5.8** Why can ′ t the perceptron-convergence rule be applied to multilayer networks?

5.3 Training Multilayer Networks

Multilayer networks are organized into different layers, each containing a number of units, which typically are not connected to each other. All networks contain an input layer, an output layer, and a number (possibly 0) of what are called *hidden layers.* The hidden layers are so called because they are connected only to other network units. They are hidden from the " outside world. "

Information enters the network via the input layer. Each unit in the input layer receives a certain degree of activation, represented numerically. Each unit in the input layer is connected to each unit in the next layer. Each connection has a weight, again representable numerically. The most common neural networks are *feedforward* networks. Activation spreads forward

through the network. There is no spread of activation between units in a given layer or backward from one layer to the previous layer.

The spread of activation through a multilayer network is illustrated in Figure 5.10 . The diagram follows the rather confusing notation standard in the neural network literature. The usual practice is to label a particular unit with the subscript i . So, we write the name of the unit as u_i . If we want to talk about an arbitrary unit from an *earlier* layer connected to u_i , we label that earlier unit with the subscript j and write the name of the unit as u_j . Just to make things as difficult as possible, when we label the weight of the connection from u_j to u_i , we use the subscript ij , with the label of the later unit coming first. So, w_{ij} is the weight of the connection that runs *from* u_j *to* u_i .

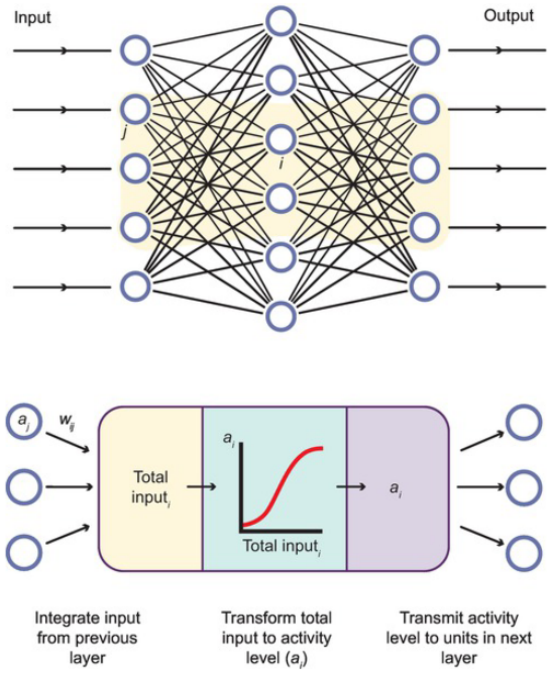


Figure 5.10 The computational operation performed by a unit in a connectionist model. Upper: General structure of a connectionist network. Lower: A closer look at unit *i*. Its operation can be broken into three steps: (1) Integrate all the inputs from the previous layer to create a total input. (2) Use an activation function to convert the total input to an activity level. (3) Output the activity level as input to units in the next layer.

(Adapted from McLeod et al. 1998 .)

As we see in the figure, our sample unit u_i integrates the activation it receives from all the units in the earlier layer to which it is connected. If there are n units connected to u_i, multiplying each by the appropriate weight and adding the resulting numbers all together gives the total input (i) to the unit. Representing the activation of each unit u_j by a_j, then we can write:

$$\text{Total input} = \sum_{j=1}^{N} w_{ij} a_j$$

The activation function then determines the unit ′ s activity level (a_i). In Figure 5.10 , the activation function is a sigmoid function. This means that a_i is low when the total input (i) is below the threshold. Once the threshold is reached, a_i increases more or less proportionally to the total input. It then levels out once the unit ′ s ceiling is reached.

Once we understand how a single unit works, it is straightforward to see how the whole network functions. It is a series of n time steps, where n is the number of layers (including the input, hidden, and output layers). In the first time step, every unit in the input layer is activated. We can write this down as an ordered series of numbers — what mathematicians call a *vector*. At step 2, the network calculates the activation level of each unit in the first hidden layer, using the process described in the previous paragraph. This gives another vector. This process continues until at step n , the network has calculated the activation level of each unit in the output layer to give the output vector.

The Backpropagation Algorithm

So, that ′ s how activation spreads through a multilayer network. But where do the weights come from? How does the network learn to solve a particular problem, whether it is computing the XOR function or distinguishing between mines and rocks (as in the network we looked at in Section 3.3)?

The basic idea of the *backpropagation algorithm* is that error is propagated backward through the network, from the output units to the hidden units. Recall the basic problem for training multilayer networks. We know what the target activation levels are for the output units. We know, for example, that a network computing XOR should output 0 when the inputs

are both 1. And we know that a mine/rock detector should output (1, 0) when its inputs correspond to a mine and (0, 1) when its inputs correspond to a rock. Given this, we can calculate the degree of error in a given output unit. But because we don ′ t know what the target activation levels are for the hidden units, we have no way of calculating the degree of error in a given hidden unit. And that seems to mean that we have no way of knowing how to adjust the weights of connections to hidden units.

The backpropagation algorithm makes it possible to calculate the error in the activation level of a given hidden unit. The basic idea is that each hidden unit connected to an output unit bears a degree of " responsibility " for the error of that output unit. If, for example, the activation level of an output unit is too low, then this can only be because insufficient activation has spread from the hidden units to which it is connected. This gives us a way of assigning error to each hidden unit. In essence, the error level of a hidden unit is a function of the extent to which it contributes to the error of the output unit to which it is connected. Once this degree of responsibility, and consequent error level, is assigned to a hidden unit, it then becomes possible to modify the weights between that unit and the output unit to decrease the error.

This method can be applied to as many levels of hidden units as there are in the network. We begin with the error levels of the output units and then assign error levels to the first layer of hidden units. This allows the network to both modify the weights between the first layer of hidden units and the output units and assign error levels to the next layer of hidden units. And so the error is *propagated* back down through the network until the input layer is reached. It is very important to remember that activation and error travel through the network in opposite directions. Activation spreads

forward through the network (at least in *feedforward* networks), whereas error is propagated backward.

How Biologically Plausible Are Neural Networks?

There are some striking dissimilarities between neural networks and the brain. For example:

- Whereas neural network units are all homogeneous, there are many different types of neurons in the brain — twelve different types in the neocortex alone.
- Brains are nowhere near as massively parallel as typical neural networks. Each cortical neuron is connected to a roughly constant number of neurons (approximately 3 percent of the neurons in the surrounding square millimeter of cortex).
- The scale of connectionist networks seems wrong. Each cortical column consists of a population of highly interconnected neurons with similar response properties. A single cortical column cuts vertically across a range of horizontal layers (*laminae*) and can contain as many as 200,000 neurons — whereas even the most complicated artificial neural networks rarely have more than 5,000 units.

There are even more striking differences when it comes to learning and training:

- Neural networks learn by modifying connection weights, and even in relatively simple networks, this requires hundreds to thousands of training cycles.

- There is no evidence that anything like the backpropagation of error takes place in the brain. Researchers have failed to find any neural connections that transmit information about error.
- Most neural networks are supervised networks and require detailed information about the extent of the error at each output unit. But very little biological learning seems to involve this sort of detailed feedback.

It is important to keep these arguments in perspective, however. For example, the principal reason it takes so long to train networks is that they tend to start with a random assignment of weights, and this is not something one would expect to find in a well-designed brain. Moreover, there are learning algorithms that are more biologically plausible than backpropagation, such as *local algorithms* .

In local learning algorithms, an individual unit ' s weight changes directly as a function of the inputs to and outputs from that unit. The Hebbian learning rule that we briefly looked at earlier is an example of a local learning rule. Neural network modelers think of it as much more biologically plausible than the backpropagation rule, and local learning algorithms are often used in networks that learn through *unsupervised* learning that does not require explicit feedback.

In *competitive networks* , for example, there is no fixed target for each output unit, and there is no external teacher. The network classifies the inputs so that each output unit fires in response to a particular set of input patterns. This works because of inhibitory connections between the output units, in contrast to standard feedforward networks, where there are typically no connections between the units in a single layer. These inhibitory

connections allow the output units to compete with each other. Each output unit inhibits the other output units in proportion to its firing rate. So, the unit that fires the most will win the competition. Only the winning unit is " rewarded " (by having its weights increased). This increase in weights makes it more likely to win the competition when the input is similar, so each output ends up firing in response to a set of similar inputs.

Competitive networks are particularly good at classification tasks. They have been used, for example, to model visual pattern recognition. One of the amazing properties of the visual system is its ability to recognize the same object from many different angles and perspectives. There are several competitive network models of this type of *position-invariant object recognition* , including the VisNet model of visual processing developed by Edmund Rolls and T. T. Milward. VisNet is designed to reproduce the flow of information through the early visual system (as sketched in Section 3.2). It has different layers intended to correspond to the stages from area V1 to the inferior temporal cortex. Each layer is itself a competitive network, learning by a version of the Hebbian rule.

In short, there are many ways of developing the basic insights in neural network models that are more biologically plausible than standard feedforward networks that require detailed feedback and a mechanism for the backpropagation of error. And in any case, neural network models should be judged by the same criteria as other mathematical models. The results of the network need to mesh reasonably closely with what is known about the large-scale behavior of the cognitive ability being modeled. So, for example, if what is being modeled is the ability to master some linguistic rule (such as the rule governing the formation of the past tense), one would expect a good model to display a learning profile similar to that generally seen in the

average language learner. In Chapter 10 , we will look at two examples of models that do seem very promising in this regard. First, though, we need to make explicit some of the general features of the neural network approach to information processing .

5.4 Information Processing in Neural Networks: Key Features

All neural networks share some very general characteristics that distinguish them from physical symbol systems.

Distributed Representations

According to the physical symbol system hypothesis, representations are distinct and identifiable components in a cognitive system. This need not be true in artificial neural networks. There are some networks for which it holds. These are called *localist* networks. What distinguishes localist networks is that each unit codes for a specific feature in the input data. We might think of the individual units as analogs of concepts. They are activated when the input has the feature that the unit encodes. The individual units work as simple feature detectors. There are many interesting things that can be done with localist networks. But the artificial neural networks that researchers have tended to find most exciting have typically been *distributed* networks rather than localist ones.

The information that a distributed network carries is not located in any specific place. A network stores information in its pattern of weights. It is the particular pattern of weights in the network that determines what output it

produces in response to particular inputs. A network learns by adjusting its weights until it settles into a particular configuration. The upshot of the learning algorithm is that the network ' s " knowledge " is distributed across the relative strengths of the connections between different units.

No Clear Distinction between Information Storage and Information Processing

According to the physical symbol system hypothesis, all information processing is rule-governed symbol manipulation. This idea only makes sense if we can distinguish, within a cognitive system, between the representations on which the rules operate and the rules themselves — just as within a logical system, we can distinguish between symbolic formulas and the rules that we use to build those symbolic formulas up into more complex formulas and to transform them.

> **Exercise 5.9** Look back at Box 4.1 and Figure 4.2 and explain how and why the distinction between rules and representations is central to the physical symbol system and language-of-thought hypotheses.

Consider how AND might be computed according to the physical symbol system hypothesis. A system for computing AND might take as its basic alphabet the symbol 0 and the symbol 1. The inputs to the system would be pairs of symbols, and the system would have built-in rules to ensure that when the input is a pair of 1s, the system outputs a 1, whereas in all other cases, it outputs a 0. What might such a rule look like?

Well, we might think about the system along the lines of a Turing machine (as illustrated in Section 1.2). In this case, the inputs would be symbols written on two squares of a tape. Assume that the head starts just to the left of the input squares. The following program will work:

Step 1 Move one square R.

Step 2 If square contains " 1, " then delete it, move one square R, and go to Step 6. Otherwise, go to step 3.

Step 3 If square contains " 0, " then delete it, move one square R, and go to step 4.

Step 4 Delete what is in square and write " 0. "

Step 5 Stop.

Step 6 If square contains " 0, " then stop. Otherwise, go to step 7.

Step 7 If square contains " 1, " then stop.

The tape ends up with a 1 on it only when the tape started out with two 1s on it. If the tape starts out with one or more 0s on it, then it will stop with a 0. The final state of the tape is reached by transforming the initial symbol structure by formal rules, exactly as required by the physical symbol system hypothesis. And the rules are completely distinct from the symbols on which they operate.

Exercise 5.10 Write a program that will compute the function XOR.

There is no comparable distinction between rules and representations in artificial neural networks. The only rules are those governing the spread of

activation values forward through the network and those governing how weights adjust. Look again at the network computing XOR in Figure 5.9 . If we input two 1s into the network (corresponding to a pair of propositions, both of which are true), then the information processing in the network proceeds in two basic stages. In the first stage, activation spreads from the input layer to the hidden layer, and both hidden units fire. In the second stage, activation spreads from the hidden units to the output unit, and the output unit fires.

The only rules that are exploited are, first, the rule for calculating the total input to a unit and, second, the rule that determines whether a unit will fire for a given total input (i.e., the activation function). But these are exactly the same rules that would be activated if the network were computing AND or OR. These " updating rules " apply to all feedforward networks of this type. What distinguishes the networks is their different patterns of weights. But a pattern of weights is not a rule or an algorithm of any kind. Rather, it is what results from the application of one rule (the learning algorithm), and it is itself the input for another rule (the updating algorithm).

The Ability to Learn from " Experience "

This capacity to learn makes neural networks a powerful tool for modeling cognitive abilities that develop and evolve over time. Unlike the AND program described earlier, they are not fixed in how they respond to inputs because they can change their weights. Supervised-learning algorithms, such as the backpropagation algorithm, change the weights in direct response to explicit feedback about how the network ' s actual output diverges from the intended output. But networks can also engage in unsupervised learning (as

we saw when we looked briefly at competitive networks). Here the network imposes its own order on the inputs it receives, typically by means of a local learning algorithm, such as some form of Hebbian learning .

Summary

Connectionist neural networks offer models of information processing very different from the physical symbol system hypothesis discussed in Chapter 4 . They are constructed from individual units that function as highly idealized neurons. In the first part of the chapter, we looked at single-layer networks and saw how they can learn via the perceptron-convergence rule. Unfortunately, single-layer networks are limited in the functions that they can compute. In contrast, multilayer networks with hidden units can compute any function that can be computed by a digital computer. The backpropagation learning algorithm makes it possible to train multilayer neural networks. The chapter ended by considering the biological plausibility of neural networks and summarizing some of the crucial differences between artificial neural networks and physical symbol systems.

Checklists

Neurally Inspired Information Processing

1. Understanding how the brain processes information requires understanding how the activities of large populations of neurons give rise to complex sensory and cognitive abilities.

2. Existing techniques for directly studying the brain do not allow us to study what happens inside populations of neurons.

3. Computational neuroscientists use mathematical models (neural networks) to study populations of neurons.

4. These neural networks are made up of units loosely based on biological neurons. Each unit is connected to other units so that activation levels can be transmitted between them as a function of the strength of the connection.

Single-Layer Networks

1. Single-layer networks can compute some Boolean functions (e.g., AND, OR, and NOT).

2. Single-layer networks can learn by adjusting their weights to minimize their degree of error (the δ signal) according to the perceptron-convergence rule.

3. Single-layer networks can only learn to compute functions that are linearly separable.

Multilayer Networks

1. Multilayer networks have hidden units that are neither input units nor output units.

2. Hidden units enable multilayer networks to learn to compute functions that cannot be learned by single-layer networks (e.g., functions that are not linearly separable).

3. The backpropagation learning algorithm for multilayer networks adjusts the weights of hidden units as a function of how " responsible " they are for the error at the output units.

Biological Plausibility

1. Neural network units are much more homogeneous than real neurons, and real neural networks are likely to be both much larger and less parallel than network models.

2. The backpropagation algorithm is not very biologically plausible. There is no evidence that error is propagated backward in the brain, and nature rarely provides feedback as detailed as the algorithm requires.

3. However, there are other learning algorithms. Competitive networks that use Hebbian learning do not require explicit feedback, and there is evidence for local learning in the brain.

Information Processing in Neural Networks

1. Representation is distributed across the units and weights, rather than being encoded in discrete symbol structures, as in the physical symbol system hypothesis.

2. There are no clear distinctions to be drawn within neural networks, either between information storage and information processing or between rules and representations.

3. Neural networks are capable of sophisticated forms of learning, which makes them particularly suitable for modeling how cognitive abilities are acquired and how they evolve.

Further Reading

The *Handbook of Brain Theory and Neural Networks* (Arbib 2003) remains the most comprehensive single-volume source for different types of computational neuroscience and neural computing, together with entries on neuroanatomy and many other " neural topics. " It contains useful introductory material and " road maps " but is somewhat out of date. Stein and Stoodley 2006 and Trappenberg 2010 are user-friendly introductions to neuroscience and computational neuroscience, respectively. Arbib 1987 surveys the theoretical issues in modeling the brain from a mathematical perspective. Glass 2016 is a neuroscience-inspired cognitive psychology textbook. Kriegeskorte and Douglas 2018 introduces the interface of cognitive science, artificial intelligence, and computational neuroscience.

The classic sources for connectionism are the two volumes of Rumelhart et al. 1986 . Churchland and Sejnowski 1992 is an early manifesto for computational neuroscience. See also Bechtel and Abrahamsen 2002 and the relevant chapters of Dawson 1998 . There are useful article-length presentations in Rumelhart 1989 (reprinted in Haugeland 1997) and Churchland 1990b (reprinted in Cummins and Cummins 2000). McLeod et al. 1998 covers both the theory of neural networks and their modeling applications, including the VisNet model of visual processing originally presented in Rolls and Milward 2000 and reviewed in Rolls 2021 . The first chapter is reprinted in Bermúdez 2006 . Dawson 2005 is a " hands-on " introduction to connectionist modeling. For a survey of applications of connectionist networks in cognitive psychology, see Houghton 2005 . See also Thomas and McClelland ' s chapter on connectionist modeling in Sun 2008 . A more recent discussion of connectionism can be found in

McClelland et al. 2010 , with commentary and target articles from others in the same issue. Connectionism went out of fashion for a few years but has recently seen a resurgence in the context of deep machine learning. See the Further Reading section in Chapter 12 for references. Hassabis et al. 2017 discusses neuroscience-inspired artificial intelligence more generally.

The biological plausibility of artificial neural networks has been much discussed, and researchers have developed a number of learning algorithms that are alternatives to backpropagation — see O ' Reilly and Munakata 2000 and Warwick 2012 . Lillicrap et al. 2020 discusses backpropagation and the brain. Bowers 2009 and Plaut and McClelland 2010 offer conflicting views on biological plausibility, as well as local and distributed representations. See Nadeau 2021 for processing similar to parallel distributed processing (PDP) in the basal ganglia and thalamus.

The perceptron-convergence rule discussed in Section 5.2 is also known as the *delta rule* . It is very closely related to the model of associative learning in classical (Pavlovian) conditioning independently developed by the psychologists Robert Rescorla and Allen Wagner in the 1970s. For reward learning and the delta rule, see Chapter 6 of Trappenberg 2010 . The *Encyclopedia of Cognitive Science* also has an entry on perceptrons (Nadel 2005). For more on McCullough and Pitts, see Chapter 2 of Arbib 1987 and Piccinini 2004 , as well as Schlatter and Aizawa 2008 . Du and Swamy 2019 is a more mathematical introduction to perceptrons.

One of the key distinguishing features of neural networks is that their " knowledge " is distributed across units and weights. This raises a number of issues, both practical and theoretical. Rogers and McClelland 2004 develops a distributed model of semantic knowledge. Philosophers have explored the relation between distributed representations and standard ways of thinking

about propositional attitudes and mental causation. Some of the points of contact are explored in Clark 1989 , 1993 . Macdonald and Macdonald 1995 collects some key papers, including an important debate between Smolensky and Fodor about the structure of connectionist networks. Other collections include Davis 1993 and Ramsey et al. 1991 .

Not all neural networks are distributed. There are also localist networks. Whereas in distributed networks, it is typically not possible to say what job an individual unit is doing (and when it is possible, it usually requires knowing a lot about what other units are doing), units in localist networks can be interpreted independently of the states of other units. For a robust defense of the localist approach, see Page 2000 and the papers in Grainger and Jacobs 1998 .

One topic not discussed in the text is the computational power of artificial neural networks. It is sometimes suggested that connectionist networks are computationally equivalent to digital computers (by virtue of being able to compute all Turing-computable functions), which might be taken to indicate that connectionist networks are simply implementations of digital computers. The implementation thesis is canvassed both by opponents of connectionism (Fodor and Pylyshyn 1988) and by leading connectionist modelers (Hinton et al. 1986). Siegelmann and Sontag 1991 present a neural network that can simulate a universal Turing machine. For skeptical discussion, see Hadley 2000 .

Chapter 6

Applying Dynamical Systems Theory to Model the Mind



Overview

6.1 Cognitive Science and Dynamical Systems

What Are Dynamical Systems?

The Dynamical Systems Hypothesis: Cognitive Science without Representations?

6.2 Applying Dynamical Systems: Two Examples from Child Development

Two Ways of Thinking about Motor Control

Dynamical Systems and the A-Not-B Error

Assessing the Dynamical Systems Approach

Overview

Pursuing the basic idea that cognition is information processing, we have explored two different approaches to information processing — the physical symbol system hypothesis and the neural networks model. Both approaches

are committed to thinking of cognition as essentially a process of transforming representational states that carry information about the agent and about the environment, although they think about these representational states in very different ways.

This chapter introduces a very different approach. Some cognitive scientists have proposed using the mathematical and conceptual tools of dynamical systems theory to model cognitive skills and abilities. We ′ ll see in Section 6.1 that dynamical systems models are importantly different from the information-processing models we have been looking at. Then, in Section 6.2 , we examine two examples of how dynamical systems models can shed light on child development.

6.1 Cognitive Science and Dynamical Systems

The dynamical systems hypothesis calls for cognitive science to be freed from its dependence on ideas of representation and computation. It holds that we can understand how organisms respond to the environment and orient themselves in it without assuming that there are internal cognitive systems that carry out specific information-processing tasks. The basic currency of cognitive science is not the information-carrying representation, and cognition does not involve computations and algorithms.

Instead, it maintains, cognitive scientists need to use the tools of dynamical systems theory to understand how perceivers and agents are embedded in their environments. Dynamical systems have been studied in physics and other natural sciences for many centuries. What ′ s new is the

idea that we can understand how cognition works by thinking of cognitive agents as dynamical systems.

What Are Dynamical Systems?

A dynamical system is any system that evolves over time in a law-governed way. The solar system is a dynamical system. So are you and I. So is a dripping tap. And so, for that matter, are Turing machines and artificial neural networks. What marks out the dynamical systems hypothesis is the idea that cognitive systems should be studied with the tools of dynamical modeling.

Dynamical modeling exploits powerful mathematical machinery to understand how certain types of natural phenomena evolve over time. Newtonian mechanics is perhaps the most famous example of dynamical modeling. All dynamical models have certain basic features.

Dynamical models typically track the evolving relationship between a relatively small number of quantities that change over time. They do this using calculus and differential or difference equations.

- *Difference equations* allow us to model the evolution of a system that changes in discrete steps. So, for example, we might use difference equations to model how the size of a biological population changes over time — each step being a year, for example.
- *Differential equations* , in contrast, allow us to model quantities that change continuously, such as the acceleration of a falling object.

One of the basic theoretical ideas in dynamical systems modeling is the idea of a *state space.* The state space of a dynamical system is a geometric way of thinking about all the possible states that a system can be in. A state space has as many different dimensions as it has quantities that vary independently of each other — as many different dimensions as there are degrees of freedom in the system, in other words. Any state of the dynamical system will involve the system having a particular value in each dimension. And so, we can uniquely identify the state of the system in terms of a particular set of coordinates in the system ' s state space.

You can think about the state space of a swinging pendulum, for example, as having two dimensions — one corresponding to its angle of displacement from the vertical and one corresponding to its angular velocity. So, every possible state that the pendulum can be in can be represented by a pair of numbers, which in turn correspond to a point in a two-dimensional space.

If we add another dimension to the state space to represent time, then we can start thinking about the evolution of the pendulum in terms of a *trajectory* through state space. A trajectory through state space is simply a sequence of points in the multidimensional space. This sequence of points represents the successive states of the pendulum.

One basic aim of dynamical systems modeling is to write equations governing the evolution of the system — that is, governing the different possible trajectories that the system can take through state space, depending on where the system starts (the system ' s *initial conditions*).

Let ' s go back to our simple pendulum. For simplicity, we ' ll assume that it is in an environment with no friction so that we won ' t have to take velocity into account. The initial condition is the pendulum ' s position at

the moment it is released and allowed to start swinging. The state space is the different possible positions that the weight can take at a given time. At any given moment, the position of the weight is fixed solely by the pendulum ′ s amplitude (its angle of displacement from its equilibrium position, which is hanging straight down) and the length of time it has been swinging.

We can represent the swinging of the pendulum as a trajectory through state space over time, as illustrated in Figure 6.1 . The state space is two-dimensional because there are two dimensions of variation. The first dimension of variation is its angular displacement from the vertical equilibrium position. The second dimension is time, which is represented vertically. So, each state of this simple dynamical system is a position – time pair — a point in the two-dimensional state space.

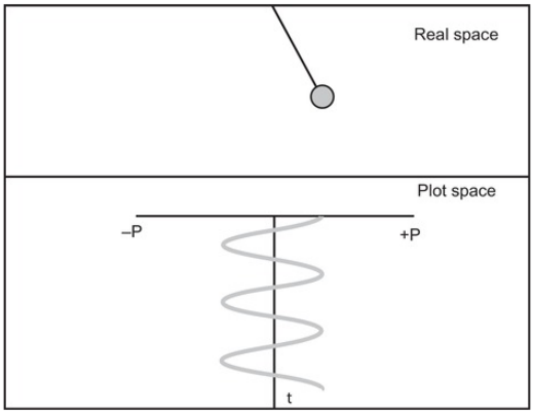


Figure 6.1 The trajectory through the state space of a swinging pendulum. The pendulum ′ s position is its displacement from the vertical (positive to the right and negative to the left). The time axis goes vertically downward.

If we remove the simplifying assumption that there is no friction, then our dynamical system becomes more complicated. Because what friction

does is decrease velocity through energy loss, we now need to add a third dimension to the state space. This third dimension represents velocity. So now, each state of the system is a position – time – velocity triple — a point in the three-dimensional state space. (See Figure 6.2 .)

Exercise 6.1 Explain in your own words the key differences between Figures 6.1 and 6.2 .

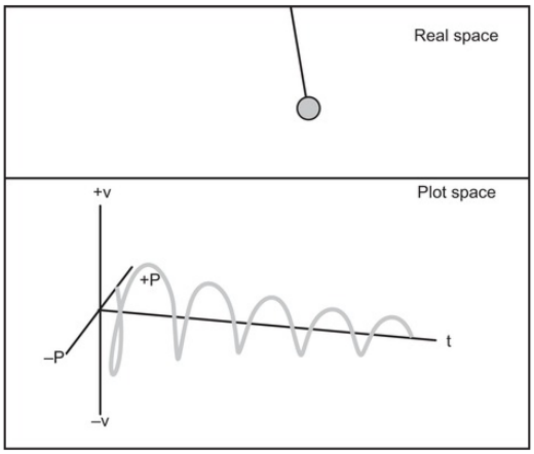


Figure 6.2 The state space of a swinging pendulum in a three-dimensional state space. This state space includes a dimension representing velocity, in order to capture energy loss and decreased velocity due to friction.

In order to model the evolution of the system, we need to be able to write equations that describe how a value on one dimension (say, its angle of displacement from the vertical) is determined by values on the other dimensions (say, velocity and time). There is no equation that will fix the position of a swinging pendulum as a function simply of velocity and time (even if we remove the effects of friction). More information is required.

To work out where the pendulum will be at a particular moment, we need to take into account two different forces. We need to factor in both the force with which the pendulum is moving and the force of gravity (which acts as a restoring force, counteracting the initial force). When the restoring force exceeds the initial force, then the pendulum starts moving back to equilibrium. And so, it continues to oscillate indefinitely (in a system where there is no friction). And in order to measure the effects of those forces, you also need to take into account the length of the pendulum.

With all this information, we can write an equation for determining the angle of displacement as a function of these other quantities. It turns out that with some important simplifying assumptions, the following equation will work when the maximum angle of displacement is relatively small (and there is no friction):

$$A = A_{\mathrm{MAX}} \sin \sqrt{(g/l)}t$$

where A is the angular displacement, A_{MAX} is the maximum angular displacement, g is the gravitational acceleration, and l is the length of the pendulum.

The equations get much more complicated when friction is brought back into the picture, and the maximum angle of displacement gets larger. But this should be enough to illustrate the basic idea of a dynamical system, and the evolution of a dynamical system can be modeled as a trajectory through state space.

The Dynamical Systems Hypothesis: Cognitive Science without

Representations?

To see how this is relevant to cognitive science, consider a famous example discussed by philosopher Tim Van Gelder, a prominent advocate for dynamical systems. Van Gelder introduces us to two ways of thinking about an engineering problem whose solution was a vital step in the Industrial Revolution.

For Van Gelder, cognitive scientists are essentially trying to reverse engineer the mind — they are trying to work out how the mind is configured to solve the problems that it deals with. Cognitive scientists have tended to tackle this *reverse-engineering* problem in a particular way — by assuming that the mind is an information-processing machine. But Van Gelder uses the example of the *Watt governor* to argue that it is neither the only way nor the best way.

The first steam engines were only capable of a reciprocating pumping motion until Scottish engineer James Watt designed a gearing system to allow steam engines to drive a flywheel and hence produce rotational power. This gearing system made it possible to use steam engines for weaving, grinding, and other industrial applications.

Unfortunately, there was still a problem. The type of applications for which steam power was needed required the power source to be as uniform as possible, which required the speed of the flywheel to be as constant as possible. But this was very hard to achieve because the speed of the flywheel depended on two things that were constantly changing — the pressure of the steam driving the engine and the amount of work that the engine was doing. What was needed (and Watt ended up inventing) was a *governor* that would regulate the speed of the flywheel .

The problem is clear, but how could it be solved? Van Gelder identifies one possible approach. This approach employs the sort of task analysis that is typical of traditional cognitive science and that is often presented in a boxes-and-arrows diagram. It breaks the task of regulating the speed of the flywheel into a series of subtasks, assumes that each of those subtasks is carried out in separate stages, and works out an algorithm for solving the problem by successively performing the subtasks. This approach gives what Van Gelder terms the *computational governor* , which follows something like the following algorithm:

1. Measure the speed of the flywheel.

2. Compare the actual speed S_1 against the desired speed S_2 .

3. If $S_1 = S_2$, return to step 1.

4. If $S_1 \neq S_2$, then:

a. Measure the current steam pressure.

b. Calculate the required alteration in steam pressure.

c. Calculate the throttle adjustment that will achieve that alteration.

5. Make the throttle adjustment.

6. Return to step 1.

The computational governor has certain features that should be very familiar by now:

It is representational. It needs some way of representing the speed of the flywheel, the pressure of the steam, and the state of the throttle valve.

It is computational . It compares, transforms, and manipulates these representations.

It is sequential . It works in a discrete, step-by-step manner.

It is decomposable (or, as Van Gelder puts it, *homuncular*). It is made up of distinct and semiautonomous subsystems, each responsible for a particular subtask — the speed measurement system, the steam measurement system, the throttle adjustment system, and so on.

So, the computational governor is an application of some of the basic principles that cognitive scientists use to understand how the mind works. The basic fact of the matter, though, is that Watt went about things in a very different way.

The Watt governor, which Watt developed using basic principles already exploited in windmills, has none of the key features of the computational governor. It does not involve representations and hence, as a consequence, cannot be computational. It is not sequential. And it is not decomposable. It is, in fact, best studied using the tools of dynamical systems modeling.

The Watt governor is illustrated at the top of Figure 6.3 . The flywheel is at the bottom. Coming up from the flywheel is a rotating spindle. The spindle rotates at a speed determined by the speed of the flywheel. It has two metal balls attached to it. As the speed of the spindle ′ s rotation increases, centrifugal force drives the metal balls upward. As the speed decreases, the balls drop down. Watt ′ s key idea was to connect the arms from which the metal balls are suspended directly to the throttle valve for the steam engine. Raising the arms closes the throttle valve, whereas the valve is opened up when the arms fall.

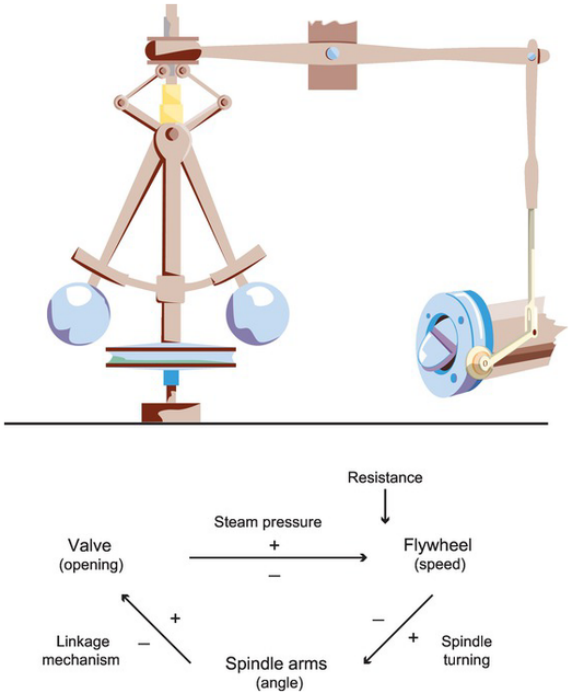


Figure 6.3 Illustration of the Watt governor, together with a schematic representation of how it works.

(From Figure 2 in Bechtel 1988.)

This ingenious arrangement allows the governor to regulate the speed by compensating almost instantaneously, whether the speed of the flywheel is overshooting or undershooting. The lower part of Figure 6.3 illustrates the feedback loop.

Van Gelder stresses four very important features of the Watt governor:

Dynamical system: It is relatively straightforward to write a differential equation that will specify how the arm angle changes as a function of the engine speed. The system is a typical dynamical system because these equations have a small number of variables.

Time sensitivity: The Watt governor is all about timing. It works because fluctuations in the speed of the flywheel are almost instantly

followed by variation in the arm angle. The differential equations governing the evolution of the system track the relation over time between the flywheel speed and the arm angle.

Coupling: The Watt governor works because the arm angle, the throttle valve, and the speed of the flywheel are all interdependent. The system as a whole is what dynamical systems theorists call a *coupled system* characterized by feedback loops.

Attractor dynamics: For any given engine speed, there is an equilibrium arm angle — an angle that will allow the engine to continue at that speed. This equilibrium arm angle is an *attractor* — a point in state space to which many different trajectories will converge. (See Box 6.1.)

So, the Watt governor can be characterized using the tools of dynamical systems theory. It is a coupled system that displays a simple version of attractor dynamics because it contains basins of attraction (as described in Box 6.1). Unlike the computational governor, it does not involve any representation, computation, or decomposable subsystems. Finally, the Watt governor works in real time. The adjustments are made almost instantaneously, exactly as required. It is very hard to see how the computational governor would achieve this.

Exercise 6.2 Explain in your own words the principal differences between the computational governor and the Watt governor.

But again, what has this got to do with the mind? It is not news, after all, that steam engines are not cognitive systems.

Van Gelder and other supporters of the dynamical systems hypothesis argue that the same basic tools that explain how the Watt governor works can be used to illuminate the workings of the mind. But the issue is not just about explanation. Dynamical systems theorists think that the explanations work because they track the basic design principles of the mind. They think not only that the mind is a dynamical system but also that when we look at the relation between the organism and the environment, what we see is a coupled system. The organism – environment complex is a system whose behavior evolves as a function of a small number of variables.

Certainly, the real test of this idea must come in concrete applications. The plausibility of the dynamical systems hypothesis cannot rest solely on an analogy between the mind and a steam engine — however suggestive that analogy may be. Some very exciting work has been done by cognitive scientists on creating dynamical systems models of particular cognitive abilities. Much of the most interesting research has been done on motor skills and motor learning. Dynamical systems theory has proved a powerful tool for understanding how children learn to walk, for example. The next section looks at two applications of the dynamical systems approach to child development.

BOX 6.1 Basins of Attraction in State Space

A particular dynamical system evolves through time along a trajectory in state space. The particular trajectory that it takes is a function of its initial conditions. So, the trajectory of the swinging pendulum, for example, is typically determined by its initial amplitude, together with the way that allowances for friction are built into the system.

But not all regions of state space are created equal. There are some regions of state space to which many different trajectories converge. These are called *basins of attraction.* In the case of a swinging pendulum subject to friction, there is a region of state space to which all trajectories converge — this is the point at which the pendulum is stationary, its equilibrium point.

Many dynamical systems have a number of basins of attraction — these are the *nonlinear dynamical systems.* There is a two-dimensional example in Figure B6.1 .

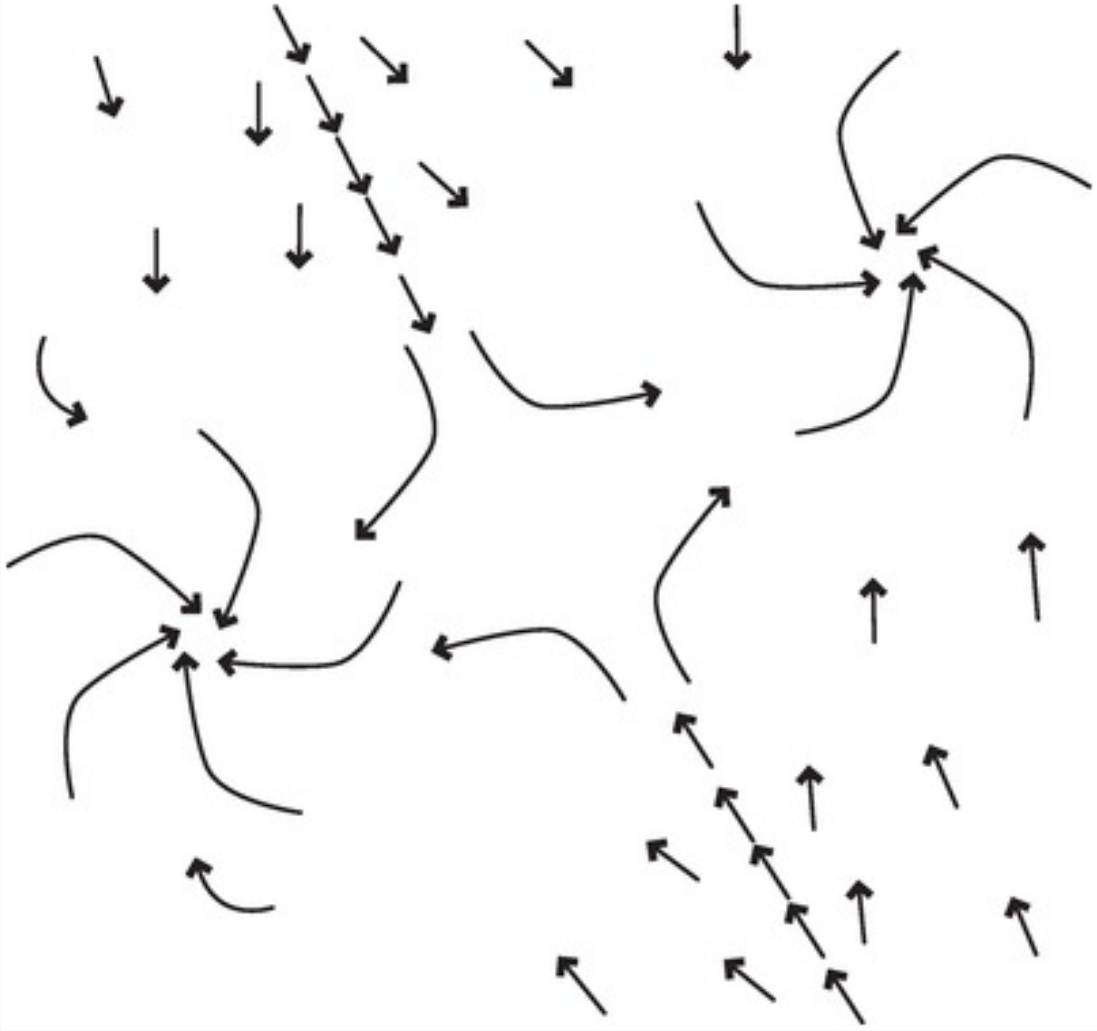


Figure B6.1

The figure illustrates a range of different possible trajectories. The trajectories are marked by arrows, with the length of the arrow indicating the speed (and hence the strength of the attraction). The state space has two basins of attraction.

Figure B6.2 gives another way of representing basins of attraction, in terms of what is often called an *energy landscape* . This offers a different way of visualizing how a system evolves through state space.

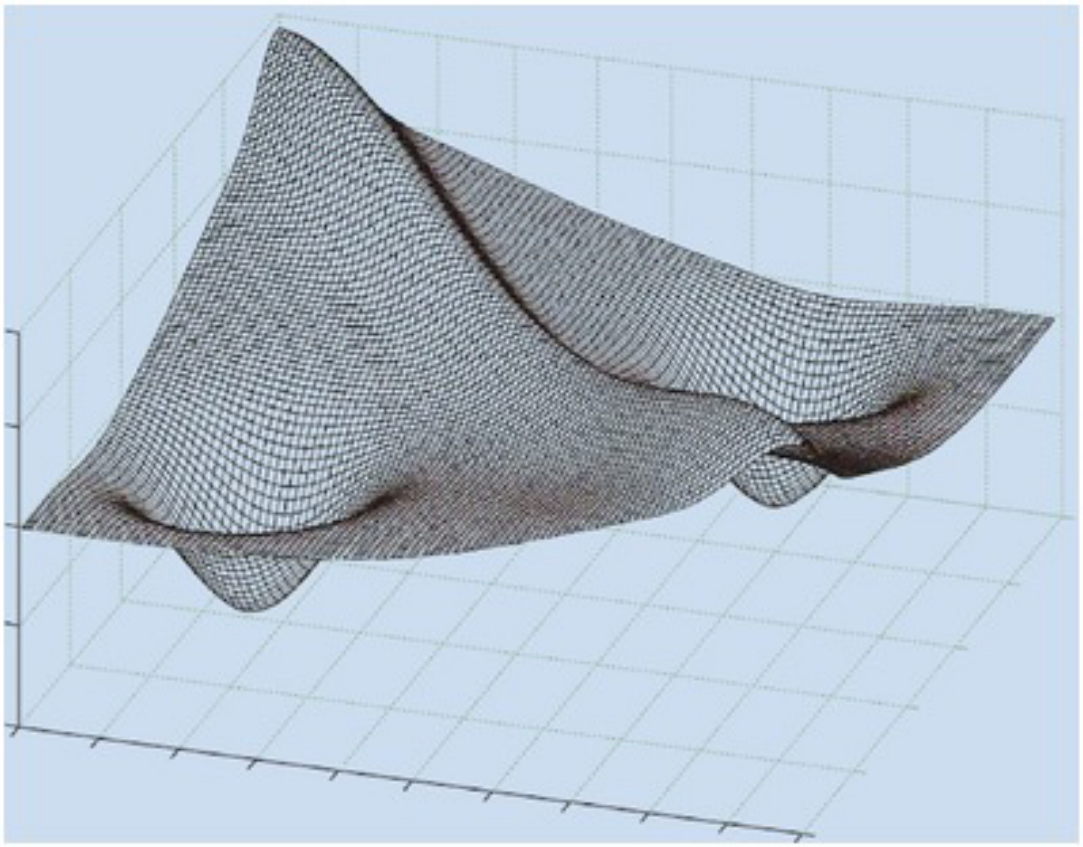


Figure B6.2

The undulating surface represents the space of possible trajectories. The two basins of attraction are represented by depressions in the surface. Because dynamical systems evolve toward a reduction in energy, trajectories will typically " roll " downhill until they end up in one of the two basins of attraction. In fact, in this particular dynamical system, any trajectory must begin on one side of the dividing " ridge " or the other — and so will end up in the corresponding basin of attraction.

6.2 Applying Dynamical Systems: Two Examples from Child Development

Dynamical models are extremely time-sensitive, that is, able to track the evolution of a system over time in very fine detail. This suggests that one profitable area to apply them is in studying how children learn new skills and abilities. In this section, we look at two concrete examples of how episodes in child development can be modeled by dynamical systems theory.

Two Ways of Thinking about Motor Control

Our first example comes from motor control, specifically how infants learn to walk. There is a direct analogy with the Watt governor. The dominant approach to understanding how movements are planned and executed is the computational model of motor control (the equivalent of the computational governor). Dynamical systems theorists offer an alternative, noncomputational way of thinking about how movements are organized and how motor skills emerge.

Consider the movement of reaching for an object. According to the computational model, planning this movement has to begin with the central nervous system calculating both the position of the target object and the position of the hand, using information both from vision and from proprioception (e.g., sensors in the arm detecting muscle flexion). It calculates a trajectory from starting position to goal position and then computes a sequence of muscle movements that will take the hand along that trajectory. Finally, executing the movement requires calculating

changes in the muscle movements to accommodate visual and proprioceptive feedback.

This gives a multistage sequence of computations that seems tailor-made for algorithmic information processing. Figure 6.4 illustrates a computational model of motor control that fits this general description. It is a standard information-processing diagram — which is often called a *boxological* diagram (because it is drawn in terms of boxes and arrows, with different cognitive tasks assigned to different boxes and the arrows indicating the flow of information processing).

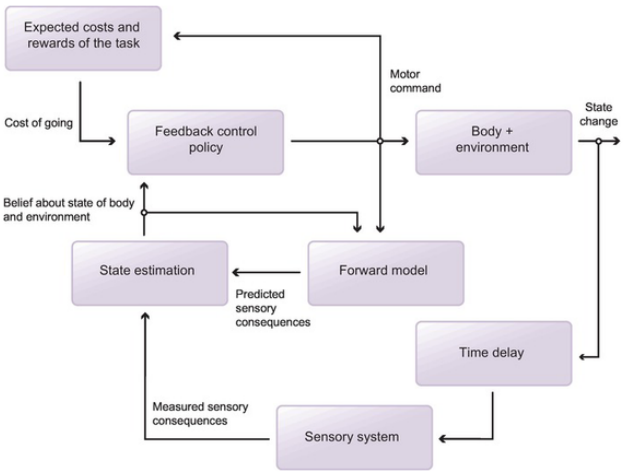


Figure 6.4 An example of the computational approach to motor control. This model incorporates both forward mechanisms (i.e., mechanisms that make predictions about the sensory consequences of particular movements) and comparator mechanisms (i.e., mechanisms that compare the predictions with actual sensory feedback).

(Adapted from Shadmehr and Krakauer 2008 .)

But psychologists Esther Thelen and Linda Smith have argued that walking is not a planned activity in the way that many cognitive scientists have assumed, following the computational approach to motor control. It

does not involve a specific set of motor commands that " program " the limbs to behave in certain ways. Instead, walking involves complex interactions and feedback loops between muscles, limbs, and different features of the environment.

Concrete evidence for Thelen and Smith ' s position comes from studies on how infants learn to walk. Most normal infants start learning to walk toward the end of their first year — at around eleven months. For the first few months, infants are capable of making stepping movements. They stop making these movements during the so-called *nonstepping window* . The movements obviously reappear when the infant starts walking.

The traditional explanation for this U-shaped developmental trajectory is that the infant ' s initial stepping movements are purely reflexive. They disappear during the nonstepping window because the cortex is maturing enough to inhibit reflex responses — but is not sufficiently mature to bring stepping movements under voluntary control.

Thelen and Smith challenged this approach experimentally. They discovered that stepping movements could be artificially induced in infants by manipulating features of the environment. So, for example, infants in the nonstepping window will make stepping movements when they are suspended in warm water. Stepping during the nonstepping window can also be induced by placing the infants on a treadmill. The treadmill increases leg strength by moving the leg backward and exploiting its spring-like properties. Stepping movements can also be inhibited before the start of the nonstepping window — for example, by attaching small weights to the baby ' s ankles.

This potential for manipulating infant stepping movements is problematic for the cortical maturation approach — because stepping

movements vary independently of how the cortex has developed. And they also point toward a dynamical systems model by identifying the crucial parameters in the development of infant walking — parameters such as leg fat, muscle strength, gravity, and inertia. The brain and the rest of the central nervous system do not have a privileged position in generating this complex behavior. Instead, we have a behavior that can, in principle, be modeled by equations tracking the interdependence of a small number of variables. Thelen and Smith have worked this idea out in great detail with a wealth of experimental studies and analyses.

Still, although walking is certainly a highly complex activity, it is not a very cognitive one. Is there support for the dynamical systems approach in a more cognitive sphere? Several examples suggest that there is. The dynamical systems approach has been profitably applied to the study of human decision-making, for example.

The Decision Field Theory developed by Jerome R. Busemeyer and James T. Townsend sets out to explain certain experimental results in behavioral economics and the psychology of reasoning in terms of the interplay of seven parameters (where agents have a choice between two actions). These parameters include settings for the strength threshold that preferences need to exceed if they are to lead to action, as well as settings for the average gain from each action. A single difference equation exploits these seven parameters to fix the agent ′s preference at a given moment.

Another example, and one that we will look at in more detail, also derives from the work of Thelen and Smith on infant development. Thelen and Smith have developed a dynamical systems approach to how young infants understand objects.

Dynamical Systems and the A-Not-B Error

Object permanence is the infant ′ s understanding that objects continue to exist when they are no longer being perceived. As we will see in much more detail in Chapter 11 (which is dedicated to object perception), object permanence emerges in stages and is intimately connected with the infant ′ s emerging " folk physics "— with the infant ′ s sensitivity to the basic principles governing how physical objects behave.

One of the first to study the development of object permanence was the famous Swiss developmental psychologist Jean Piaget. In his highly influential 1954 book *The Construction of Reality in the Child* , Piaget described a very interesting phenomenon.

One index of infants ′ understanding of object permanence is whether and how they search for hidden objects. Up to around seven months, infants are very poor at searching for objects even when they have just seen them being hidden. For the very young infant, out of sight seems to be, quite literally, out of mind. From twelve months or so onward, infants search normally. But between the ages of seven months and twelve months, young infants make a striking error that Piaget termed the *stage IV error* and that is now generally known as the *A-not-B error* . Figure 6.5 illustrates a typical experiment eliciting this error.

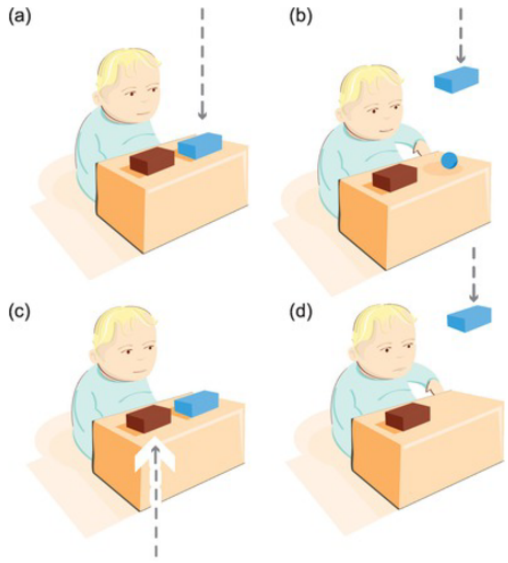


Figure 6.5 The stage IV search task, which typically gives rise to the A-not-B error in infants at around the age of nine months. (a) The experimenter hides an object in the left-hand box. (b) The infant searches successfully. (c) But when the experimenter moves the object in full view of the infant, (d) the infant searches again at the original location.

(Adapted from Bremner 1994 .)

Infants are placed in front of two containers — A and B. They see a toy hidden in container A and reach for the toy repeatedly until they are habituated to its presence in container A. Then, in plain view, the experimenter hides the toy in container B. If there is a short delay between hiding and when the infants are allowed to reach, they will typically reach toward container A rather than toward container B (even though they have just seen the toy hidden in container B).

Piaget himself explained the A-not-B error in representational terms. He suggested that it is not until they are about twelve months old that infants are able to form abstract mental representations of objects. Before then, their actions are driven by sensorimotor routines. In the first stage of

the task, searching for the toy in container A allows the infant to discover the spatial relationship between the toy and the container. But this knowledge only exists in the form of a sensorimotor routine. It cannot be extrapolated and applied to the new location of the toy. And so, infants simply repeat the routine behavior of reaching toward container A.

Exercise 6.3 Give Piaget ′ s explanation of the A-not-B error in your own words.

Other cognitive and neural interpretations have been proposed. In one common interpretation, the key factor is the infant ′ s ability to inhibit the reaching response to container A. The first part of the task effectively conditions the infant to make a certain response (reaching for container A), and it is only when the infant becomes able to override that response that she can act on her knowledge of where the toy is. This ability to inhibit responses is tied to the maturation of the prefrontal cortex, which is generally held to play an important role in the executive control of behavior.

However, as with infant stepping movements, it turns out that infant performance on the task can be easily manipulated. It is well known, for example, that the effect disappears if the infants are allowed to search immediately after the toy is hidden in container B. But Smith, Thelen, and other developmental psychologists produced a cluster of experiments in the 1990s identifying other parameters that had a significant effect on performance:

- Drawing infants ′ attention to the correct side of their visual field (e.g., by tapping on a board on the same side as the toy)

significantly improves performance. Directing their attention the other way has the opposite effect.

- The most reliable predictor of infant performance is the number of times the infants reach for the toy in the preliminary A trials.
- The error can be made to disappear by changing the infant ′ s posture — eight-month-old infants who are sitting during the initial A trials and then supported in a standing position for the B test perform at the same level as twelve-month-old infants (see Figure 6.6).



Figure 6.6 An infant sitting for an A trial (*left*) and standing for a B trial (*right*). This change in posture causes younger infants to search as twelve-month-old infants do.

(From Smith and Thelen 2003 .)

If the A-not-B error were primarily a cognitive phenomenon, due either to the infants ′ impoverished representational repertoire or their undeveloped cortical executive system, then it should not be so variable and easy to manipulate. It is hard to think of a cognitive/neural explanation for why standing up should make such a drastic difference, for example.

As in the infant walking case, Smith, Thelen, and their collaborators propose a dynamical systems model — the *dynamic field model.* The dynamic field represents the space in front of the infant — the infant ′ s

visual and reaching space. High levels of activation at a specific point in the dynamic field are required for the infant to reach toward that point. Thelen and Smith think about this in terms of a threshold. Movement occurs when the activation level at a particular point in the dynamic field is higher than the threshold.

The dynamic field model traces the evolution of activation levels in the dynamic field over time as a function of three different types of input:

- *Environmental input:* This parameter represents the constraints the environment poses on the infant ' s possible actions. It will vary, for example, according to whether the infant is sitting or standing. The environmental input parameters also include the attractiveness and salience of the target, as well as its distance, and contextual features of the environment, such as visual landmarks.
- *Task-specific input:* This reflects the specific demands placed on the infant — the experimenter drawing attention to the target, for example.
- *Memory input:* The strength of this input is a function of the infant ' s previous reaching behavior. Because reaching behavior is partly a function of environmental input and task-specific input, the memory input reflects the history of these two types of input. And as one would expect, it is weighted by a decay function that reflects how time diminishes memory strength.

All of these parameters are coded in the same way, in terms of locations in the movement/ visual field. This allows them all to contribute to raising the

activation level above the threshold for a specific location (either container A or container B).

And this, according to Smith and Thelen, is exactly what happens in the A-not-B error. The perseverative reaching takes place, they claim, when the strength of the memory input overwhelms the other two inputs. This is illustrated in Figure 6.7 .

Exercise 6.4 Explain in your own words how the dynamic field model differs from computational accounts of the A-not-B error.

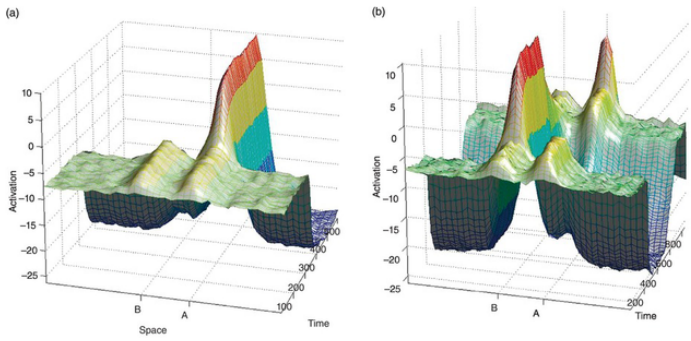


Figure 6.7 Applying the dynamic field model to the A-not-B error. (a) The time evolution of activation in the planning field on the first A trial. The activation rises as the object is hidden, and because of self-organizing properties in the field, it is sustained during the delay. (b) The time evolution of activation in the planning field on the first B trial. There is heightened activation at A before the hiding event, owing to memory for prior reaches. As the object is hidden at B, activation rises at B, but as this transient event ends, because of the memory properties of the field, activation at A declines, and that at B rises.

Their explanation makes no general appeal to cortical maturation, executive control, or the infant ′ s representational capacities. And it is very sensitive to how the initial conditions are specified. If the strength of the memory input is allowed to diminish (by increasing the delay before the infant is allowed to reach, for example), then one would expect the error to diminish correspondingly — as indeed happens. The same holds for the other experimental manipulations that Smith and Thelen have uncovered. These manipulations all subtly change the inputs and parameters in the model, resulting in changes in the activation levels and hence in the infant ′ s reaching behavior.

Assessing the Dynamical Systems Approach

The experiments and models produced by Smith, Thelen, and other dynamical systems theorists clearly give us powerful tools for studying the evolution of cognition and behavior. The explanations that they provide of the A-not-B error and how infants learn to walk are simultaneously more complex and simpler than the standard type of information-processing explanations current in cognitive science. They seem more complex because they bring a wide range of factors into play that cognitive scientists had not previously taken into account, and they steer us away from explanations framed in terms of a single information-processing mechanism and toward complex systems with subtle interdependencies and time sensitivity. At the same time, their explanations seem simpler because they do not invoke representations and computations.

We started out, though, with the idea that the dynamical systems approach might be a radical alternative to some of the basic assumptions of

cognitive science — and in particular to the idea that cognition essentially involves computation and information processing. Some proponents of the dynamical systems approach have certainly made some very strong claims in this direction. Van Gelder, for example, has suggested that the dynamical systems model will, in time, completely supplant computational models, and thus traditional cognitive science will end up looking as quaint (and as fundamentally misconceived) as the computational governor.

But claims such as these ignore one of the most basic and important features of cognitive science. Cognitive science is both interdisciplinary and multilevel. The mind is too complex a phenomenon to be fully understood through a single discipline or at a single level. This applies to the dynamical systems hypothesis no less than to anything else. There is no more chance of gaining a complete picture of the mind through dynamical systems theory than there is of gaining a complete account through neurobiology, say, or artificial intelligence (AI). All of these disciplines and approaches give us deep, but partial, insights.

The contrast that Van Gelder draws between the computational governor and the Watt governor is striking and thought-provoking, but it cannot be straightforwardly transferred from engineering to cognitive science. The computational governor and the Watt governor do seem to be mutually exclusive. If we are trying to solve that particular engineering problem, we need to take one approach or the other — but not both. Nothing like this holds when it comes to cognition, however. Dynamical systems models are perfectly compatible with information-processing models of cognition.

Dynamical systems models operate at a higher level of abstraction. They allow cognitive scientists to abstract away from details of

information-processing mechanisms in order to study how systems evolve over time. But even when we have a model of how a cognitive system evolves over time, we will still need an account of what makes it possible for the system to evolve in those ways.

Here is an analogy. Dynamical systems theory can be applied in all sorts of areas. So, for example, traffic jams have been modeled as dynamical systems. Physicists have constructed models of traffic jams that depend on seeing traffic jams as the result of interactions between particles in a many-particle system. These models have proved surprisingly effective at predicting phenomena such as stop-and-go traffic and the basic fact that traffic jams often occur before a road ' s capacity has been reached.

This certainly gives us a new way of thinking about traffic, as well as new predictive tools that make it easier to design roads and intersections. But no one would ever seriously propose that this new way of thinking about the collective movement of vehicles means that we no longer have to think about internal combustion engines, gasoline, spark plugs, and so on. Treating a traffic jam as an effect in a multiparticle system allows us to see patterns that we couldn ' t see before. This is because it gives us a set of tools for abstracting away from the physical machinery of individual vehicles. But " abstracting away from " is not the same as " replacing. " Cars can be modeled as particles in a multiparticle system — but these models only make sense because we know that what is being modeled are physical objects powered (by and large) by internal combustion engines.

With this analogy in mind, look again at the dynamical field model in Figure 6.7 . This model may well accurately predict the occurrence of the A-not-B error in young infants. But look at what it leaves out. It says nothing about how memory works, how the infant plans her movement,

how she picks up the experimenter ′ s cues, and so on. We don ′ t need answers to these questions in order to construct a dynamical system model. But neither can we simply leave them unanswered. The dynamical systems approach adds a powerful tool to the cognitive scientist ′ s toolkit, but it is unlikely ever to be the only tool .

Summary

This chapter has explored how the mathematical and conceptual tools of dynamical systems theory can be used to model cognitive skills and abilities. These models plot how a system evolves over time as a function of changes in a small number of system variables. We looked at two examples of dynamical systems models of child development. Dynamical systems theory offers fresh and distinctive perspectives both on how infants learn to walk and on infants ′ expectations about objects that they are no longer perceiving (as revealed in the so-called A-not-B error). Despite the more radical claims of some dynamical systems theorists, however, it is unlikely that traditional information-processing models will completely disappear from cognitive science.

Checklists

Some cognitive scientists have turned to dynamical systems theory as an alternative to traditional information-processing models of cognition.

1. The dynamical systems approach in cognitive science proposes studying cognitive systems with the tools of dynamical systems theory.

2. Dynamical models use calculus-based methods to track the evolving relationship between a small number of variables over time — a trajectory through state space.

3. Dynamical systems often display coupling (interdependencies between variables) and an attractor dynamics (there are points in the system ′s state space on which many different trajectories converge).

4. Cognitive systems modeled using dynamical systems theory do not display many of the classic features of information-processing systems. Dynamical models typically are not representational, computational, sequential, or homuncular.

Dynamical systems theory permits time-sensitive models of learning and skill acquisition in children.

1. Case studies include learning to walk in infancy and performance on the A-not-B search task.

2. Support for the dynamical systems approach comes from experiments showing that performance can be drastically altered by manipulating factors that would typically be ignored by computational models.

3. The explanatory power of the dynamical systems approach does not mean that it should *replace* information-processing approaches to cognitive science.

4. The dynamical systems approach sheds light on cognitive systems at a particular level of organization. There is no reason to think that the level of explanation it provides should be the only one in cognitive science.

Further Reading

Timothy Van Gelder has promoted the dynamical systems approach to cognitive science in Van Gelder 1995 and 1998 . The papers in Port and Van Gelder ′ s *Mind and Motion: Explorations in the Dynamics of Cognition* (1995) contain influential studies and models (including Townsend and Busemeyer ′ s model of decision-making), as well as theoretical statements. Thelen and Smith ′ s 1993 edited volume *A Dynamical Systems Approach to the Development of Cognition and Action* provides more detail on their studies of infant walking, as well as contributions from other dynamical systems theorists. Their *Behavioral and Brain Sciences* article (Thelen et al. 2001) presents the model of the A-not-B error (with extensive peer commentary). Smith and Thelen 2003 is a more accessible introduction. Jacobson and Degotardi 2022 is a dynamical systems account of infant – caregiver joint attention.

The January 2012 issue of the journal *Topics in Cognitive Science* is devoted to the complex systems approach to cognitive science, which is a branch of dynamical systems theory. For an application of the dynamical

systems approach to different areas of cognitive psychology, see Spivey 2007 . For overviews and assessments of the dynamical systems approach to cognitive science, see Eliasmith 1996 , Clark 1998 , Clark 2001 : Chapter 7, Weiskopf 2004 , Clearfield et al. 2009 , Spencer et al. 2009 , Needham and Libertus 2011 , Spencer et al. 2011 , Riley and Holden 2012 , and Spencer et al. 2012 . For more recent reviews, see Samuelson et al. 2015 and Perone and Simmering 2017 (and Favela 2020 for an annotated bibliography). Favela 2021 focuses on dynamical systems theory in neuroscience, whereas Constant et al. 2021 argues that predictive processing models can accommodate both representational and dynamicist approaches.

Chapter 7

Bayesianism in Cognitive Science



Overview

7.1 Bayesianism: A Primer

Degrees of Belief and Subjective Probability

Conditional Probability

Bayes ′ s Rule (the Short Version)

7.2 Perception as a Bayesian Problem

The Predictive Challenge of Perception

Case Study: Binocular Rivalry

7.3 Neuroeconomics: Bayes in the Brain

What Is Expected Utility?

Case Study: Neurons That Code for Expected Utility

Overview

Bayesianism has become increasingly important in cognitive science. It offers a (relatively) simple model of problem solving and decision-making that has proved very profitable for modeling the mind. This chapter introduces the basic

principles of Bayesianism, illustrating them through two very different case studies.

As we see in Section 7.1 , the basic idea of Bayesianism is that an organism ' s information about the world is modeled as an assignment of probabilities to different propositions about what is going on in the world. Bayesian probabilities are not like the probability that a fair coin will fall heads nor like life-expectancy tables calculated from huge studies of mortality rates. Those are objective probabilities, based on frequencies. Bayesian probabilities, in contrast, are subjective. They reflect an organism ' s best guess, which is updated as new information comes in, according to a (relatively) simple rule called Bayes ' s rule.

Section 7.2 shows how perception can be seen as a Bayesian problem. Perceptual systems have to work backward from noisy data to a consistent, coherent, and — it is hoped — accurate model of how the world is. This is an inference problem. The brain has to make a reverse inference from the noisy data to a hypothesis about the layout of the world. This is exactly the kind of inference for which Bayes ' s rule is particularly well suited. We will illustrate this through a Bayesian model of the phenomenon of *binocular rivalry* .

The theory of expected utility is an extension of Bayesian principles to decision-making. Section 7.3 introduces the principle of expected utility and explains how expected utility is calculated. We then look at an example of Bayesian modeling from *neuroeconomics* , which uses economic tools such as the theory of expected utility to study the brain.

7.1 Bayesianism: A Primer

Remarkably, given how powerful it is as a modeling tool, the basic elements of Bayesianism are really pretty straightforward. There are three key ideas:

- Belief comes in degrees.
- Degrees of belief can be modeled as probabilities, and so they have to conform to the basic principles of the probability calculus.
- Learning takes place by updating probabilities according to Bayes ' s rule.

This section presents these three key ideas in turn.

Degrees of Belief and Subjective Probability

It is natural to talk about cognitive systems having beliefs about their environment. But what are beliefs? Most people think about beliefs as on/off states. You either believe something or you do not. Either you believe that there is a predator hiding in the woods, or you do not. Correspondingly, your belief is either true or false. There is no in-between state of partially believing something.

Bayesians reject this approach. They think that belief comes in degrees. At one end of the spectrum are things in which you are completely confident. At the other end are things in which you have absolutely no confidence in at all. From a Bayesian perspective, the interesting things all happen in between these two extremes. Most of what we believe about the world we have some confidence in — but not complete confidence.



Figure 7.1 An illustration purporting to be of Thomas Bayes from a 1936 book on the history of life insurance.

(From Wikimedia Commons.)

> **Exercise 7.1** Give examples of (a) something in which you have complete confidence, (b) something in which you have no confidence whatsoever, (c) something in which you are fairly confident but not completely so, and (d) something in which you have a slight degree of confidence.

The second key idea of Bayesianism is that degrees of belief are probability assignments. Bayesians replace the vague idea of having more or less confidence in some proposition with the much more precise notion of assigning a particular numerical probability to that proposition. So, rather than describing an organism as being more confident than not, but still some way short of completely confident, that there is a predator in the woods, a Bayesian might say that the organism assigns a probability of 0.7 to there being a predator in the woods.

Exercise 7.2 Go back to Exercise 7.1 and assign numerical probabilities to your answers to (a) through (d).

It is essential to Bayesians that degrees of belief obey the fundamental principles of the probability calculus. So, we can use the rules of the probability calculus to update and combine our degrees of belief. Box 7.1 offers a quick refresher on the basic rules of probability, but here are some examples that will probably seem familiar:

- If we assign probability p to some sentence S, then we have to assign probability $1 - p$ to **not-S** , the negation of S.
- If we assign probability p to S and probability q to R, and we know that S and R are (probabilistically) independent of each other, then we have to assign probability $p \times q$ to **S AND R** (i.e., to S and R both holding).
- If we assign probability p to S and probability q to R, and we know that S and R are mutually exclusive, then we have to assign probability $p + q$ to **S OR R** (i.e., to at least one of S and R holding).

BOX 7.1 Basic of the Probability Calculus

There are different but equivalent ways of presenting the probability calculus. Probabilities are sometimes assigned to sentences and sometimes to events. If you think in terms of sentences, then it is most natural to formulate the probability calculus in logical terms (talking about conjunctions and disjunctions, for example). If you think in terms of events, then it is most natural to use elementary set theory (talking about intersections and unions of events, for example). I ′ ll use the framework of sentences.

There are four basic principles defining how probabilities behave:

Basic principle 1: Probabilities are numbers between 0 and 1.

Basic principle 2: All impossible sentences have probability 0.

Basic principle 3: All necessary truths (such as " 2 + 2 = 4 ") have probability 1.

Basic principle 4: If sentences P and Q are logically equivalent, then p (P) = p (Q).

Then, with these basic principles in place, the probability calculus contains simple rules that tell us how to assign probability to complex sentences built up from simpler sentences for which we know the probabilities.

The Negation Rule

If sentence S has probability p , then its negation **not-S** (¬S) has probability 1 − p .

The Disjunction Rule (Restricted)

If sentences R and S are mutually exclusive, then the probability of **R or S** is p (R) + p (S).

The Conjunction Rule (Restricted)

If sentences R and S are independent of each other (i.e., the presence of one does not make the other more likely), then the probability of **R and S** is p (R) × p (S).

In order to lift the restrictions, we need to apply the concept of **conditional probability** , as explained in the text. The conditional probability of S conditional upon R, written p (S/R), is the probability that S holds, on the assumption that R holds. So, for example, the probability of throwing a 4 with a six-sided die is 1/6. But the probability of throwing a 4, conditional upon throwing an even number, is 1/3.

We can use conditional probability to define the conjunction rule.

The Conjunction Rule (General)

$$p(\text{R and S}) = p(\text{R/S}) \times p(\text{S}).$$

And then, with this general definition of conjunction, we can define the disjunction rule.

The Disjunction Rule (General)

$$P(\text{R or S}) = p(\text{R}) + p(\text{S}) - p(\text{R and S}).$$

That ′ s it!

The most important inference rule for reasoning with degrees of belief is Bayes ′ s rule. But to understand Bayes ′ s rule, we first need to understand the concept of conditional probability.

Conditional Probability

Here is the basic idea. The probability of A, conditional upon B, is the probability that A holds, relative to the assumption that B holds. In other words, you assume for the moment that B holds and then calculate the probability of A on that assumption. We write the probability of A, conditional upon B, like this:

$$p(\mathrm{A}/\mathrm{B})$$

So, if " A " stands for " There is thunder " and " B " for " It is raining, " then *p* (A/B) is the probability that there is thunder, if we assume that it is raining.

Exercise 7.3 Explain why if " A " stands for " There is thunder " and " B " stands for " It is raining, " then you would expect *p* (A/B) to be greater than *p* (A).

You can think about probabilities in terms of a sample space. The sample space is the full range of possible situations. Individual probabilities are regions of the sample space, which we measure by how much of the sample space they occupy. So, for example, in Figure 7.2 , the probability of A is the proportion of the sample space in which A is true. If A is true in 30 percent of

the sample space, then the probability of A is 0.3. And the same holds for B, of course.

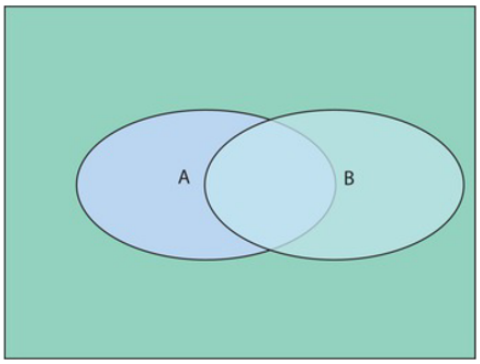


Figure 7.2 A diagram showing (a) the proportion of the probability space in which A is true, (b) the proportion of the probability space in which B is true, and (c) the intersection of A and B (which is the region where A and B are both true).

Now, what about the probability of the conjunction A&B? Well, it ′ s easy to see what that is from the diagram. The probability *p* (A&B) is the proportion of the sample space in which both A and B are true — where the A space overlaps with the B space.

So, finally, what about the probability of A conditional upon B? Well, for *p* (A/B), we also need to consider the region of the sample space that is both A and B. This is the intersection of A and B. But we are not looking at how that region of the sample space relates to the sample space as a whole. Because we are looking at A, conditional upon B, we are interested only in the relation between the A and B region of the space and the B region of the space. In other words, *p* (A/B) is the proportion of the intersection of A and B to the B space.

Well, if you find that convincing, then you will have no problem with the formal definition of conditional probability because the formula for conditional probability is basically a translation of what I ′ ve just said into the language of probability. Here it is:

$$p(\mathrm{A/B}) = \frac{p(\mathrm{A\&B})}{p(\mathrm{B})}.$$

In other words, to derive the conditional probability *p* (A/B), you take the probability of A and B both holding (i.e., *p* [A&B]) and divide by the probability that B holds.

Exercise 7.4 Suppose you have a fair coin, which you toss twice. Let " A " stand for " The first toss comes up heads " and " B " for " The second toss comes up heads. " Calculate (a) *p* (A), (b) *p* (B), and (c) *p* (A&B). Then (d) use the formula in the text to calculate *p* (A/B) — that is, the probability that the second toss comes up heads, conditional on the first toss coming up heads .

Bayes ' s Rule (the Short Version)

To see what Bayes ' s rule allows us to do, it helps to think about situations in which you have a hypothesis that you are trying to decide whether to accept or reject. You have some evidence for the hypothesis. So, the obvious question to ask is: How strong is that evidence?

When we ask that question, what we are really asking is: How likely is it that the hypothesis is true, given the evidence? And that, in turn, is really a question about conditional probability. We are asking about the probability of the hypothesis conditional upon the evidence: *p* (Hypothesis/Evidence). Using obvious abbreviations, what we are trying to discover is *p* (H/E).

Often, in this sort of situation, we have information about how likely the evidence is, given the hypothesis. In other words, we might know the

conditional probability *p* (E/H). In that situation, what we need to do is to find a way of " inverting " this probability so that we can get from *p* (E/H) to *p* (H/E).

Imagine a medical scenario. You test positive for a nasty disease. So, the hypothesis (H) is that you actually have the disease. The evidence (E) is the positive test. And suppose you know how reliable the test is. Then you know how likely it is that you will test positive if you actually do have the disease. In other words, you know *p* (E/H). But what you really want to know is *p* (H/E) — that is, how likely it is that you have the disease, given that you have tested positive.

Here is some terminology that tends to appear in discussions of Bayesian approaches to cognitive science.

Posterior Probability

This is the probability that you end up with after applying Bayes ' s rule. It is *p* (H/E).

Prior Probability

This is the probability that you originally assign to the hypothesis. It is *p* (H).

Likelihood of the Evidence

This is the probability that you ' ll get the evidence, if the hypothesis is true. It is *p* (E/H).

To continue with the medical example, the likelihood of the evidence is the reliability of the test. If it accurately detects 99 percent of cases of the disease, then the likelihood of the evidence is 0.99. The prior probability would

be given by the frequency of the disease in the population. So, if the disease afflicts 1 in 10,000 people, then the prior probability would be 0.0001.

So, using this terminology, we can write Bayes ′ s rule in words like this:

$$\textbf{Posterior probability of the hypothesis} = \frac{\textbf{Likelihood of the evidence} \times \textbf{Prior probability of the hypothesis}}{\textbf{Probability of the evidence}}.$$

If you remember this formulation in words, then you ′ ll understand the basic conceptual foundation of Bayesian updating. If you can also remember the formula in the probability calculus, as given in Box 7.2 , then you will be able to plug numbers in and apply Bayes ′ s rule to solve specific problems.

BOX 7.2 Deriving Bayes ′ s Rule

Here ′ s the formula for Bayes ′ s rule:

$$p(\mathrm{H}/\mathrm{E}) = \frac{p(\mathrm{E}/\mathrm{H})p(\mathrm{H})}{p(\mathrm{E}/\mathrm{H})p(\mathrm{H}) + p(\mathrm{E}/\neg\mathrm{H})p(\neg\mathrm{H})}.$$

The top of the equation (the numerator) is the probability of the evidence, conditional upon the hypothesis, multiplied by the probability of the hypothesis. This amounts to the probability of the evidence and the hypothesis both holding. In other words, *p* (E/H) *p* (H) = *p* (E&H).

The bottom of the equation (the denominator) has two parts, connected by an addition sign. The first part is the same as the numerator. It is essentially *p* (E&H). So, it is the probability of the evidence and the hypothesis both holding. The second part works the same way, so *p* (E/¬H) *p* (¬H) comes out as *p* (E&¬H). This is the probability that the evidence holds but not the hypothesis. In other words, that you have a false positive.

So, the denominator gives the probability that **either** you have a true positive test (the hypothesis and the evidence both hold) **or** you have a false positive test (where the evidence holds but not the hypothesis).

But these are the only two options. Because you ′ ve had a positive test, it has to be either true or false. So, what the denominator really adds up to, then, is simply the probability of getting a positive test in the first place. In other words, you can write the denominator much more simply as *p* (E).

Putting this all together, then, Bayes ′ s rule simplifies to

$$p(H/E) = \frac{p(E\&H)}{p(E)}.$$

But now — look where we have arrived. This is a straightforward application of the definition of conditional probability.

Exercise 7.5 To make sure that you understand the reasoning here, go through the reverse of the simplification process, starting with the definition of conditional probability and ending up with Bayes ′ s rule.

Bayesian agents update their beliefs by applying Bayes ′ s rule. If you are a Bayesian agent, then you start off with a set of prior probabilities. These are the probabilities that you assign to the different hypotheses about how things might turn out. You are also aware of how likely it is, for each of these hypotheses, that you will encounter different forms of evidence. So, you know various likelihood probabilities. And then, as the evidence comes in, you apply Bayes ′ s rule to derive your posterior probabilities.

It ′ s important to remember all of the different things that go into Bayes ′ s rule. It is easy, for example, to underestimate the importance of the priors. But that can be a grave mistake, as you can see by looking at the equation again. The lower the prior probability of the hypothesis is, then the lower the numerator will be in the equation (i.e., the part above the line). Because the denominator (the part below the line) is less than 1, that means that the posterior probability will end up being lower.

This means that if the prior probability is very low, then the posterior will also be very low, no matter what the other values are. This is very important in the context of medical diagnosis. Suppose again that you test positive for a nasty disease that is extremely rare. Even if the test is highly reliable (say, 0.9999), then it is highly unlikely that you actually have the disease.

Here ′ s why. Suppose that only 1 person in every 100,000 has the disease. Now imagine that 100,000 are tested. The test is highly reliable, so we can assume that that person will test positive. But even though the test is highly reliable, it will still misdiagnose 1 person in every 10,000. So, there will be around 10 false positives in our population of 100,000 people. That means that there will be 11 people testing positive out of 100,000, only one of whom actually has the disease. So, your odds are actually pretty good!

Exercise 7.6 Work through the example in the text, but instead of the prior probability being 1 in 10,000, take it to be 1 in 100. How likely is it now that you have the disease?

7.2 Perception as a Bayesian Problem

After working through the theory and the math, it is time to apply it. To see how Bayes ′ s rule might be used in cognitive science modeling, we will consider visual perception — and in particular, a puzzling phenomenon known as *binocular rivalry* .

The Predictive Challenge of Perception

Perception is an obvious place to apply Bayesian ideas. Our perceptual systems deliver a model of the environment that is, by and large, fairly accurate — at least for practical purposes. But this model is grossly underdetermined by the information that actually reaches the sensory systems. So, how do our perceptual systems get from proximal sensory stimulation to a full-blown model of the distal environment?

Back in the second half of the nineteenth century, German physicist and psychological pioneer Hermann von Helmholtz suggested that perception was a process of unconscious inference. This idea was taken up in the New Look perceptual psychology pioneered by Jerome Bruner in the 1940s. Some decades later, in 1983, Irving Rock published a famous book entitled *The Logic of Perception* , which offered a very distinctive version of the inferential model of perception.

Rock was very much influenced by the Gestalt school of perceptual psychology, including figures such as Max Wertheimer, Kurt Koffka, and Wolfgang Köhler (in Germany) and Rudolf Arnheim (in the USA). He took from them the idea that perceptual inference is largely a top-down and holistic process. From the Gestalt perspective, the visual system imposes structure on an essentially unstructured retinal image, using general principles about how objects and groups of objects form organized patterns. These are the famous Gestalt principles of grouping. Four of these principles are illustrated in Figure 7.3 .

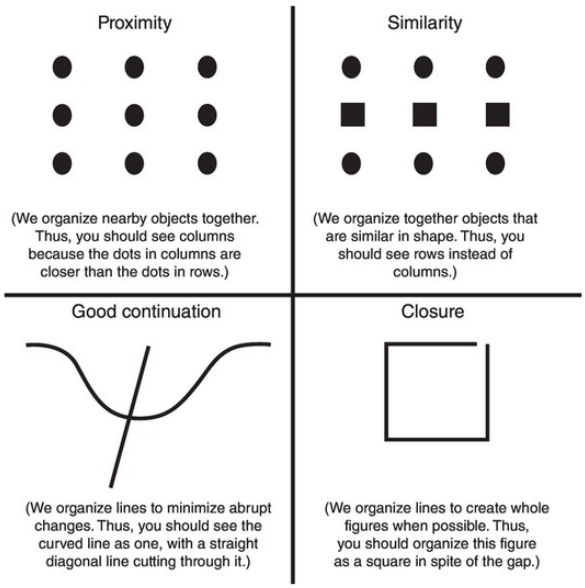


Figure 7.3 Four of the seven Gestalt principles of grouping, illustrated and explained.

(From www.skidmore.edu/~hfoley/PercLabs/Shape.htm .)

But it is reasonable to ask exactly how those Gestalt principles work. How do those general principles of organization and grouping structure how we see the world?

Bayesian models of perception have got an answer. Perceptual systems make probabilistic inferences, exploiting a constantly updated body of probabilistic knowledge. This probabilistic knowledge can be described in terms of the basic Bayesian concepts introduced in the last section:

- Perceptual systems aim to derive a hypothesis (H) about the layout of the distal environment.
- All they have to go on is the evidence (E) provided by sensory stimulation at the retina, or at the membrane window of the cochlea.
- Each perceptual system aims for the hypothesis that is most probable given the evidence — that is, conditional probabilities of the *p* (H/E)

variety. These are posterior probabilities.

- Perceptual systems store information about the likelihood of different environmental setups. These are prior probabilities — p (H).
- Perceptual systems also store information about how likely different types of sensory stimulation are, given different layouts of the distal environment. These are likelihoods, conditional probabilities of the form p (E/H).

Perhaps you can see where the Gestalt principles might fit into this overall picture?

The Gestalt principles are essentially principles about the probable structure of the environment — that is, principles about the probability of different types of environment setups. They function as Bayesian priors. Look at the principle of continuation in Figure 7.3 to appreciate this. Effectively, what this says is that a hypothesis about the layout of objects in the environment that contains abrupt changes should have a lower prior probability than one with fewer or no abrupt changes.

So, the next question is: How do perceptual systems make probabilistic inferences that end up in a model of how things are in the external environment? There are no prizes for guessing that Bayesians think that this is all done using Bayes ′ s rule!

Case Study: Binocular Rivalry

Italian scientist and playwright Giambattista della Porta discovered *binocular rivalry* in the late sixteenth century. When separate images are presented to each eye, what you actually perceive alternates between the two images. The image that you see switches seemingly at random. Figure 7.4 illustrates two examples of stimuli that can be used to elicit the phenomenon. You can check

the effect out for yourself. There are experimental demonstrations on the internet. See the Further Reading section and website for details and directions.

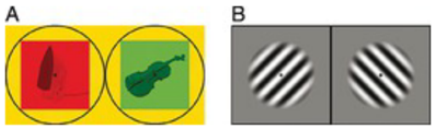


Figure 7.4 Two examples of stimuli used to elicit binocular rivalry.

(From Figure 1 in Freyberg et al. 2015 .)

Binocular rivalry is a special case of perceptual rivalry. If you look at an ambiguous figure, such as one of the examples in Figure 7.5 , then your visual system will lock onto one of the available interpretations, and you have to work hard to start seeing it the other way — to switch from seeing a duck to seeing a rabbit, for example.

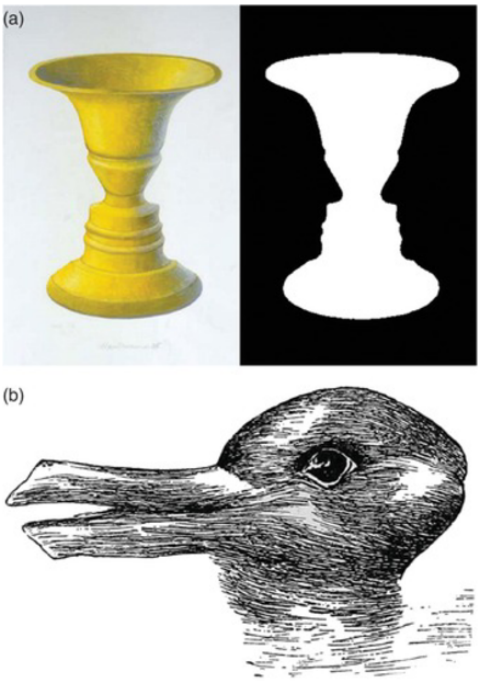


Figure 7.5 Two well-known ambiguous figures: (a) Rubin ′ s vase and (b) the duck – rabbit illusion.

(From https://upload.wikimedia.org/wikipedia/commons/b/b5/Rubin2.jpg and https://commons.wikimedia.org/wiki/File:Duck-Rabbit_illusion.jpg , respectively.)

Perceptual rivalry and binocular rivalry both involve changes in perception, even though the stimulus remains the same. In binocular rivalry, there does not seem to be anything about the image in front of either eye that prompts the switch. So, the explanation for the alternating perceptions must lie either within the visual system or downstream in more central processing.

Different explanations have been proposed for binocular rivalry. Some early theorists, including della Porta himself, took it as evidence for the view that we only ever see with one eye at a time. Others speculated that the switch might be due to attentional factors. Much more recently, though, a simple and elegant Bayesian explanation has been proposed by Jakob Hohwy, Andreas

Roepstorff, and Karl Friston in an article published in the journal *Cognition* in 2008.

Think about the challenge that the visual system faces in a typical binocular rivalry situation, like that depicted in the left-hand pair of stimuli in Figure 7.4 . The visual system has to decide what it is looking at. Does the distal environment contain a red iron? Or a green violin? Or some sort of mishmash composite object that is part iron and part violin, colored both red and green? (To simplify, I will gloss over the fact that we are really dealing with a picture of a red iron rather than a red iron.)

In Bayesian terms, there are three different hypotheses:

Hypothesis 1 (H1): Red iron

Hypothesis 2 (H2): Green violin

Hypothesis 3 (H3): Red – green iron – violin composite object

So — the visual system ' s job is to decide which hypothesis to accept, given the evidence it has. And the evidence is the images that are presented to each eye. In other words:

Evidence (E): Picture of red iron (left [L] eye) and picture of green violin (right [R] eye)

Translating this all back into the vocabulary that we looked at in Section 7.1 , what the visual system needs to figure out is which of the following conditional probabilities is the highest:

$$p(\mathrm{H1}/\mathrm{E})$$
$$p(\mathrm{H2}/\mathrm{E})$$
$$p(\mathrm{H3}/\mathrm{E})$$

These are the *posterior probabilities* — the probability, for each hypothesis, that it is true, given the available evidence. Each can be derived using Bayes ' s rule. A Bayesian visual system will accept the hypothesis with the highest posterior probability.

To apply Bayes ' s rule, the visual system also needs to know the *likelihoods*. These are conditional probabilities, specifying how likely the relevant evidence would be if the corresponding hypothesis were true — so, for example, how likely it would be to have a red iron image on the left and a green violin image on the right if what was out there was a red iron. There are three to consider:

$$p(\mathrm{E/H1})$$
$$p(\mathrm{E/H2})$$
$$p(\mathrm{E/H3})$$

The visual system also needs to consider the *prior probability* of each hypothesis. So, that gives us three more probabilities to factor in:

$$p(\mathrm{H1})$$
$$p(\mathrm{H2})$$
$$p(\mathrm{H3})$$

Bayes ' s rule tells us how to use the priors and the likelihoods to calculate the posterior probabilities.

It is important that you can apply Bayesian techniques and models to the visual system without assuming that the visual system assigns specific numbers to the priors and likelihoods. To see how this works, let ' s look again at the formula for Bayes ' s rule:

$$\textbf{Posterior probability of the hypothesis} = \frac{\textbf{Likelihood of the evidence} \times \textbf{Prior probability of the hypothesis}}{\textbf{Probability of the evidence}}.$$

The visual system is trying to compare the three hypotheses: H1, H2, and H3. The denominator (the bottom of the equation) is going to be the same for each hypothesis. (If you want to check this, look at how *p* [E] is calculated in Box 7.2 .)

So, for the purposes of comparing H1 through H3, the visual system can effectively ignore the denominator. It only needs to calculate, for each hypothesis, the product of the likelihood of the evidence and the prior probability. For H1, for example, it would need to multiply the likelihood *p* (E/H1) by the prior *p* (H1). And so on for H2 and H3.

But now suppose we make the simplifying assumption that all the likelihoods are the same. That means that the visual system thinks that the evidence (an image of a red iron on the left and a green violin on the right) would be equally likely whether there was a red iron, a green violin, or a red – green iron – violin in front of it. Then the visual system can ignore the likelihoods, too, and all it needs to look at are the prior probabilities.

Looking at the priors, it seems reasonable to think that there is no particular reason for the visual system to think that it is any more or any less probable that there would be a red iron in front of it than a green violin. So, you might think that in this case,

$$p(\mathrm{H1}) = p(\mathrm{H2}).$$

But at the same time, the visual system will probably think it very unlikely that it will run into a red – green iron – violin. So, it would assign a much

lower prior probability to H3. Hence:

$$p(\mathrm{H3}) < p(\mathrm{H1}) \quad \text{and} \quad p(\mathrm{H3}) < p(\mathrm{H2}).$$

So, we can put this all together to get the key idea in Hohwy et al.'s 2008 paper. The end result is that there are joint winners. The visual system assigns the same posterior probabilities to H1 and to H2, with both clearly preferred to H3. That gives

$$p(\mathrm{H1}/\mathrm{E}) = p(\mathrm{H2}/\mathrm{E}).$$

But now the visual system has nothing to go on to choose between H1 and H2. It could be a red iron, or it could be a green violin. So, what does it do?

Well, unable to decide between H1 and H2, the visual system switches between them, more or less at random. It doesn't try to construct some sort of composite perceptual image because that would correspond to a much less likely outcome. Put another way, the prior probability of the environment containing a composite object is much lower than the prior probability that it contains an iron or the prior probability that it contains a violin. According to Hohwy et al., the characteristic effect of binocular rivalry is a rational response to Bayesian updating.

Exercise 7.7 Write down in your own words a summary of the Bayesian explanation of binocular rivalry.

7.3 Neuroeconomics: Bayes in the Brain

A second application of Bayesian principles to study the brain comes from neuroeconomics, which is an interdisciplinary area within cognitive science. It is located at the interface between neuroscience, on the one hand, and microeconomics, the psychology of reasoning, behavioral finance, and decision theory, on the other. In brief, neuroeconomics studies how brains deal with money, investments, and risky choices. It uses the tools of neuroscience, obviously, but combines them with experimental paradigms that psychologists have developed to study reasoning and data from financial markets, as well as theoretical models from decision theory and elsewhere.

The dominant theoretical model in neuroeconomics is Bayesian expected utility theory. This is a development of the basic Bayesian approach developed and discussed in earlier sections. Bayesian expected utility theory also has the theory of probability at its core, but it goes beyond the theoretical to incorporate practical reasoning. To appreciate the distinction, consider the difference between these two questions:

What should I believe? (theoretical reasoning)

What should I do? (practical reasoning)

Probability theory and Bayes ′ s rule are enough to answer the theoretical question (according to strict Bayesians). But we need more machinery for the practical question.

That ′ s where expected-utility theory comes into play. The concept of utility allows us to model how much an individual values different outcomes. Being able to measure value in this way allows us to choose between different courses of action. We need, first, to take into account, for each available course of action, how much we value each of its different possible outcomes. And

then, second, we need to consider how likely each of those outcomes is. We ′ ll look in more detail at how this works next.

What Is Expected Utility?

Although the actual concept of utility did not appear until the nineteenth century, it has its roots in the sixteenth and seventeenth centuries, when mathematicians (and professional gamblers) were starting to develop ideas about probability in order to understand games of chance.

The best place to start is with the concept of *expected monetary value* . Here ′ s a simple example. What is the most that you would pay for the opportunity to play a game in which you win $10 if a fair coin is tossed and lands heads but get nothing if it comes up tails?

Assume that you are *risk-neutral* . You have no views either way about taking risks. You are not prepared to pay extra because you enjoy the thrill of gambling. And nor do you require any special compensation to bet money on the toss of a coin.

Intuitively, a risk-neutral person would pay up to $5 for the chance to play this game. There are two possible outcomes — heads or tails. If the coin comes up heads, she will walk away with $10. If it comes up tails, she will get $0. So, her average return will be $5. Therefore, the *expected monetary value* of this game is $5.

If you know the expected value of a gamble and your opponents do not, then you can set the odds so that you are guaranteed to come out on top over the long term. In essence, this is why casinos and lotteries always win in the long run. They set the odds so that punters are always paying more than the expected value of the game or the ticket.

But in some instances, the concept of expected value gives strange and implausible results. Suppose, for example, that you have the opportunity to play the following game.

St. Petersburg Game

> A fair coin is tossed. If it lands heads, you receive $2 and the coin is tossed again. If it lands tails, the game ends. But if it lands heads on the second toss, then you will receive $4 and the coin is tossed again. The game will continue as long as the coin keeps landing heads. At each round, the payoff (for heads) will be twice what it was in the previous round.

Think about the expected value of the St. Petersburg game. The expected value of the game is the expected value of continuing to get heads. On each toss of a fair coin, the probability of heads is 0.5. So, the expected value of the game is $(0.5 \times \$2) + (0.5 \times \$4) + (0.5 \times \$8) + \ldots$ and so on indefinitely.

You should be able to see that the expected value is infinite (if not, try rewriting the sum as $1 + 2 + 4 + 8 + \ldots$). So, if we think about this the same way as the earlier example, you should be prepared to pay an infinite amount of money to play this game. This is obviously absurd. Very few people would be prepared to pay more than a few dollars for the chance to play the St. Petersburg game. So, something must be wrong with the concept of expected value. But what?

Exercise 7.8 How much would you pay for a chance to play the St. Petersburg game? Explain your answer.

The St. Petersburg game was first proposed by the eighteenth-century Swiss mathematician Nicholas Bernoulli. His brother Daniel came up with a solution (which he published in the *Commentaries of the Imperial Academy of Sciences in St. Petersburg* — hence the name). In essence, Daniel Bernoulli concluded that we should not try to value a gamble in terms of its expected value. In its place, he introduced the concept that we now call *utility* and proposed, in effect, that we think in terms of expected utility rather than expected value.

Utility is supposed to measure your preference. If you assign more utility to X than to Y, then you prefer X to Y, and so, given the choice, you will choose X over Y. The point of introducing the concept of utility is that the utility you assign to something is not directly proportional to its expected (monetary) value.

Money has what is often called *diminishing marginal utility* . The additional utility you get from an extra $10 depends on how much money you already have. The $10 that takes your net worth from $100 to $110 will probably mean much more to you than the $10 that takes you from $1,000,000 to $1,000,010.

Combining the concept of utility with the theory of probability discussed earlier gives the Bayesian approach to practical reasoning and decision-making. The central notion is the idea of *expected utility* .

To calculate the expected utility of a possible action, start by identifying the different possible outcomes that might result from that action, and assign a utility to each of those outcomes. You need to take into account not just how much you value the different outcomes but also how likely each is. After assigning a probability to each outcome, then, for each outcome, you multiply its utility by its probability. Adding together the results of this operation for each outcome gives you the expected utility for the action.

Exercise 7.9 You are considering two possible actions — going for a swim in the river and going to watch a movie. The weather is doubtful, and there is a 40 percent chance of rain. In general, you prefer swimming to moviegoing — but not when it rains. Swimming offers 10 units of utility (standardly called *utils*) when it is not raining but only 3 utils in the rain. Going to the movies will give you 6 units of utility, irrespective of the weather. As a Bayesian decision-maker, concerned only with maximizing expected utility, which action should you choose?

This seems a long way from cognitive science, you might think. What has this all got to do with how we think about cognition? Well, in the next section, we will see how this model can be applied in a very unexpected way — to shed light on the behavior of neurons in a region of the parietal cortex.

Case Study: Neurons That Code for Expected Utility

The experiments discussed in this section all involve recording the activity of single neurons in monkeys, using a technology developed in the 1950s by Herbert Jasper at the Montreal Neurological Institute in Canada and Edward Evarts at the National Institutes of Health in the United States. Tiny electrodes are inserted into the monkey ′ s brain, which can be done while the animal is awake because there are no pain or touch receptors in the brain. The electrodes are small and sensitive enough to detect the firing rates of individual neurons.

Paul Glimcher (of New York University) and Michael Platt (from the University of Pennsylvania) aimed to examine what goes on in the parietal cortex when monkeys react to visual stimuli. They used fruit juice rewards to

train their monkeys to make specific responses to visual cues. Although it was not what they were originally looking for, they ended up finding intriguing evidence of neurons that have significant Bayesian characteristics.

To understand the experiments, you need to know a little about how the muscles around the eye work. An important part of what they do is move the eyes in order to compensate for our own movements. This allows us to have (relatively) stable perceptions of the world around us. They also move the eyes so that the high-resolution part of the eye (the *fovea*) is focused on interesting and important things in the environment. These gaze-alignment movements come in two varieties:

Saccadic eye movements move the line of sight very quickly from one place to another in the visual environment. This allows the perceiver to scan the environment (to detect a predator, for example).

Smooth-pursuit eye movements allow the eyes to track objects moving continuously in a single direction.

Monkeys can be trained to perform saccadic eye movements. In a typical experiment, they are placed in front of a screen, fixating on a colored light directly in front of them. The experimenters flash a red spot on the right side of the screen. When the monkey makes a saccadic eye movement toward the red spot, it is rewarded with a portion of fruit juice. The Glimcher and Platt experiments are variations on this basic theme.

What happens in the brain in between the monkey detecting a red spot on the right side of the visual field and its making a saccadic eye movement (a saccade) toward the red spot? Both ends of the process are relatively well understood. The primate visual system has been comprehensively mapped out, so there is no mystery about how the monkey detects the red spot.

The other end of the process is how saccades are actually generated. This is also well understood. It is known that the *superior colliculus* , which is located in the midbrain, and the *frontal eye field* , which is in the frontal cortex, both play an important role in controlling saccades. It is also widely accepted that these two brain areas are organized topographically. That means that they are organized like a map, with individual neurons responsible for specific locations to which a saccade might be directed. So, just before the monkey makes a saccade to a specific location, the neuron corresponding to that location fires in the superior colliculus and/or frontal eye field.

But what happens between detecting the red spot and initiating the saccade? This is where the lateral intraparietal area (LIP) comes in. Because the LIP projects to both the superior colliculus and the frontal eye field, it is the obvious place to look. Figure 7.6 illustrates the route of saccade production, showing why the LIP is so important.

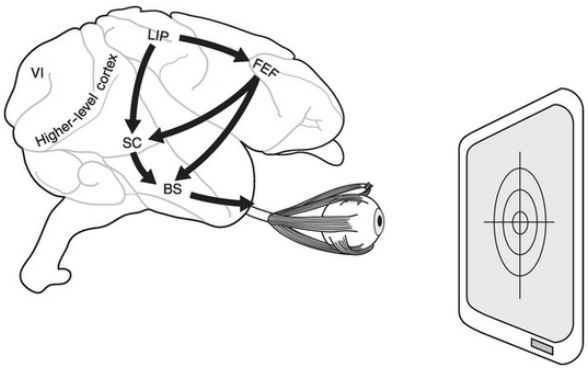


Figure 7.6 The principal pathways for saccade production. LIP = lateral intraparietal area; FEF = frontal eye field; SC = superior colliculus; BS = brain stem. Note that whereas the LIP and FEF are cortical structures, the SC and BS are much more primitive areas and not part of the forebrain at all.

(From Figure 10.2 in Glimcher 2003 : 230.)

Platt and Glimcher came up with an ingenious set of experiments suggesting that the LIP is essentially carrying out Bayesian calculations. The experiments were fairly complicated, but monkeys are extremely talented at learning things — and also highly motivated by fruit juice.

Experiments in this area are typically set up so that a particular action always leads to the same response. So, if a monkey is being trained to push a lever when it sees a red light on the right-hand side of its visual field, the correct action will always be rewarded (by delivery of fruit juice, typically). Moreover, the rewards are constant. If the reward for the correct action is 4 mL of juice on the first trial, it will still be 4 mL of juice on the twentieth trial.

Translating into our Bayesian language, what this means is that the reward is always delivered with probability 1 while the utility of the reward remains constant. Platt and Glimcher ′s breakthrough idea was to vary both probability and utility. Varying the size of the reward and how likely it is to be delivered allowed them to explore whether neurons are sensitive to those variations in the reward. In effect, it allowed them to test for Bayesian neurons.

Probability-Detecting Neurons

To explore whether neurons in the LIP are sensitive to probability, Platt and Glimcher set up a saccade experiment in which the probability that the saccade would be rewarded varied.

As usual, the monkeys started by fixating on a light at the center of the screen. Shortly afterward, two other lights appeared on the screen, one on the left of the fixation point and one on the right. These were the targets for the saccades. Then the fixation light changed color, turning either red or green. The monkeys had been trained that when the fixation light turned red, they would

be rewarded if they made a saccade to the left — while a saccade to the right would be rewarded if the light was green.

The twist to the experiment was that the fixation light was set up to turn red 80 percent of the time and green 20 percent (or the other way around, depending on the block). The experimental setup is illustrated in Figure 7.7 .

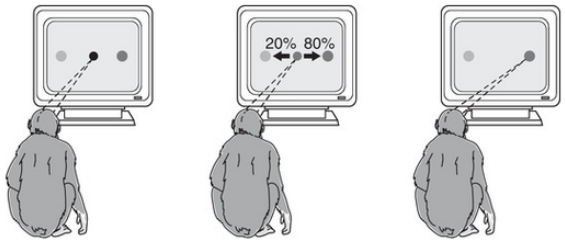


Figure 7.7 Platt and Glimcher ′ s probabilistic cued-saccade task.

(From Figure 10.11 in Glimcher 2003 : 257 .)

Platt and Glimcher assumed that over the course of a block of one hundred trials, the monkeys would have typically figured out that, say, a saccade to the left was much more likely to be rewarded than one to the right. So, they recorded throughout each trial from representative neurons in the LIP.

To make sense of the results, they just compared, for each neuron, the trials where the stimulus and the movement were the same. So, for example, they compared all the trials in a given block where the monkey made a saccade to the left in response to a red light. Doing that made it possible to compare blocks where the probability of a red light (and hence a reward for making a saccade to the left) was low with blocks where red had a high probability.

Here ′ s what they found (toward the end of the blocks, when the monkey should have learned which response was more likely to the rewarded):

- Prior to the fixation light changing color, in blocks where the probability of being rewarded with a left saccade was high, a typical LIP neuron had a much higher firing rate than in blocks where the

probability of reward was low. Note that this is the firing rate *before* the monkey receives any indication as to which direction will be rewarded.

- After the fixation light changed color, the firing rate for the saccade that was definitely going to be rewarded was maximal, even if the probability that the saccade would be rewarded had been low. Note that this is the firing rate *after* the fixation light has indicated which direction will be rewarded.

Translating this back into Bayesian terms, the first finding seems to show that the neurons are encoding the prior probabilities — the probability that a given saccade will be rewarded before it is known which saccade will be rewarded. And the second seems to show that the neurons are updating the priors to posteriors when the fixation light changes color. Once the fixation light has turned either red or green, it is certain which saccade will be rewarded, no matter whether it was a high-probability saccade or a low-probability one. And the neurons respond by firing at full blast, as it were. Figure 7.8 illustrates both findings.

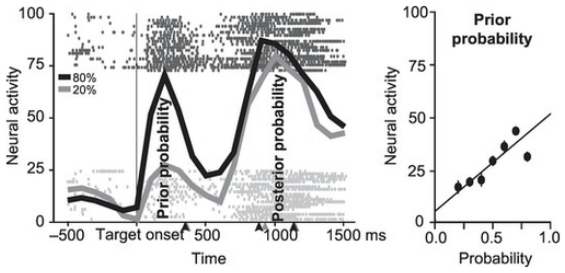


Figure 7.8 Activity of an LIP neuron during the probability experiment. Rows of tick marks in the panel indicate precise times of neural action potentials during each of twenty trials used to compute the averages, shown as thick lines.

(From Figure 10.12 in Glimcher 2003 : 260.)

You can see how the neuron ' s firing rates in the low-probability and high-probability conditions are very different before the moment when the fixation point changes color. Then they subsequently end up firing at the same rates because the changing color of the fixation point indicates that they actually will be rewarded (remember that all the trials illustrated ended with rewards).

Exercise 7.10 Explain in your own words why Platt and Glimcher think that they have discovered evidence of neurons that code for probability.

Utility-Detecting Neurons

So, what happens if probabilities (and all the other dimensions of the experiment) are held constant, but the quantity of the reward is varied? Well, as Platt and Glimcher observed, that would be a very good test of whether LIP neurons are sensitive to utility. And to explore it, they used a cued-saccade task very similar to the one we have been considering.

As before, the monkeys fixate on a point in the center of the screen. When the fixation point changes color, that indicates to the monkey that the reward will be delivered if the monkey makes a saccade to the left (red) or the right (green). The probability of red versus green was held constant, with each color coming up on exactly half the trials. What varied was the quantity of the reward. In one block, for example, the reward for looking left might be double that for looking right.

The results are illustrated in Figure 7.9 . As before, what the figure shows is the average response profiles over time of a single neuron in two different

blocks. In both blocks, the stimulus and response are held constant. They are all cases, say, where the light turns green and the monkey is rewarded for a saccade to the right. What varies is the quantity of the reward. In one block, marked by a light-colored line, the response receives a low reward. In the other, marked with a bold line, the response receives a high reward.

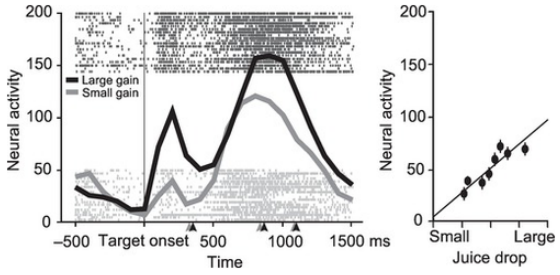


Figure 7.9 Platt and Glimcher ′ s cued-saccade experiment, with stimulus and response held constant and the quantity of reward varied. The neuron ′ s average firing rate is shown for a high-reward block (*bold line*) and a low-reward block (*light-colored line*).

(From Figure 10.13 in Glimcher 2003 : 262.)

Before the light changes color, the neuron fires more strongly on average in the high-reward block than in the low-reward block. After the light changes color (and so when the reward is revealed), the average firing rate increases significantly in both conditions. But the difference across the two conditions remains constant. The neuron responds more vigorously to the larger reward.

Strictly speaking (as Platt and Glimcher note), this experiment does not show that LIP neurons are sensitive to utility. Their firing rates correlate with the quantity of reward, but that is not necessarily the same as the utility that the monkey might derive from the reward. The experiments don ′ t reveal any analog to the phenomenon of diminishing marginal utility, for example. But still, they are highly suggestive — and exactly what one would expect in a Bayesian brain!

Combining Probability and Utility

So, the two different cued-saccade experiments just reviewed seem to uncover neurons in the LIP that code, first, for the probability of a reward and, second, for the size of a reward (and perhaps for its utility). We are close to having all the ingredients for a fully Bayesian brain, in which neurons code for something very close to expected utility — because, as we saw earlier, expected utility is really just a combination of probability and utility. The expected utility of an action is arrived at by summing the utility of its different possible outcomes, each weighted by the probability with which it will occur. That takes us to the final experiments for this case study because Platt and Glimcher came up with intriguing results here, too — not quite expected utility but intriguingly close.

To study sensitivity to expected utility, Platt and Glimcher needed to introduce an element of choice because expected utility only applies when monkeys (or people) have more than one available action. The principle of expected utility is a tool for choosing between two or more available actions (choose the action with the highest expected utility).

So, they turned the cued-saccade task into a free-choice task. As before, the monkey fixated on a spot at the center of the screen. But this time, it could choose to look at a stimulus on the left or a stimulus on the right. Everything was held constant within each block of one hundred trials, but the quantity of the (fruit juice) reward varied across blocks — 0.2 mL for looking left and 0.1 mL for looking right, for example.

From a strict Bayesian perspective, there is an optimal way of responding in this type of experiment. You should sample the two alternatives until you have worked out which yields the highest reward, then stick with that option until the end of the trial. That would be the best way to maximize expected utility over the long run.

It turns out, though, that the monkeys did not adopt the optimal strategy. Instead, they displayed a form of *matching behavior*. That is, they split their choices between the two alternatives in a way that matched the distribution of the total reward across the two alternatives. In other words, if looking right yielded twice the reward of looking left, then they looked right twice as often as they looked left.

This is not the behavior that you would expect from an animal maximizing expected utility, but Platt and Glimcher noticed something very interesting when they looked at what the monkeys were doing on a trial-by-trial basis. They saw that there was a pattern to the individual " choices " that the monkeys made. When they looked at each individual choice, they saw that the monkeys seemed to be engaged in a form of maximizing behavior. The monkeys were not maximizing expected utility, but they seemed to be maximizing something else that was not that far off from expected utility.

They used a version of *melioration theory* , developed by the animal behaviorist Richard Herrnstein (who, incidentally, first discovered the matching phenomenon when studying pigeon pecking behavior), in order to estimate how much value the monkey attached to each option in each trial. So, they were able to estimate the expected value to the monkey of looking left versus looking right on trial 47, for example.

According to melioration theory, value is fixed by local rates of reward. An animal behaving according to melioration theory will choose the option that seems most attractive at that time. And attractiveness at a time is fixed by the rewards that that option has yielded in recent trials. So, if looking left has had a good track record in yielding the reward on the past few trials, then the monkey will continue to look left. But if looking left starts to look unsuccessful, the monkey will switch and start to look right.

It turned out that the monkeys were actually behaving consistently with melioration theory, when each trial was viewed in the context of the previous ten trials. What that means is that the monkeys tended to choose the option that had been most highly rewarded over the previous ten trials. So, they were maximizing expected value, as calculated by melioration theory. This is not the same as the expected utility of each of the two options because it is a local quantity that looks back to the history of rewards, whereas expected utility is more of a global measure of overall rewards. But still, it is certainly a related measure (and in fact, expected value, in the melioration theory sense, is a good approximator of expected utility in many contexts). This is illustrated in Figure 7.10 .

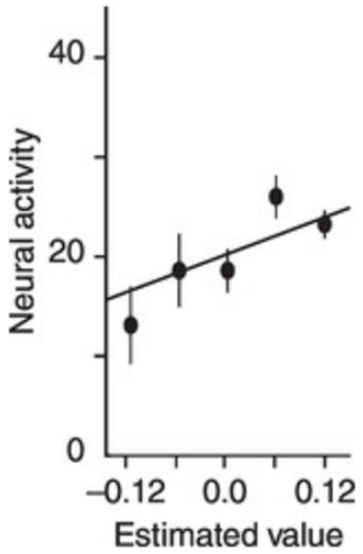


Figure 7.10 Activity of an LIP neuron while a monkey makes its own choice compared to a behaviorally derived estimate of the value of the movement to the monkey.

(From Figure 10.15 in Glimcher 2003 : 266.)

The important point from the perspective of neural economics, though, is that when Platt and Glimcher looked at the data from this perspective, they

discovered a close correlation between the firing rates of individual neurons and the anticipated value to the animal, as computed by melioration theory. So, when the estimated value to the animal of a particular option was low (based on the results of the previous ten trials), the firing rates would be low. And the greater the estimated value, the higher the firing rate. The neurons, it seemed, were firing in accordance with expected value, as calculated by melioration theory.

In conclusion, the experiments do not show that neurons in the LIP are calculators of expected utility. But they do seem to uncover neurons sensitive to the varying benefits to the animal of different courses of action. Platt and Glimcher can have the last word:

> In our free-choice task, both monkeys and posterior parietal neurons behaved as if they had knowledge of the gains associated with different actions. These findings support the hypothesis that the variables that have been identified by economists, psychologists and ecologists as important in decision-making are represented in the nervous system .
>
> (Platt and Glimcher 1999 : 266)

Summary

This chapter began by introducing the basic tools for understanding Bayesian approaches to cognitive science. Reviewing the elements of the probability calculus allowed us to introduce Bayes ′ s rule as a tool for measuring the strength of support that evidence provides for a hypothesis. We saw how this toolkit can be applied to yield a Bayesian model of binocular rivalry.

The second part of the chapter explored Bayesian approaches to choice behavior and practical decision-making. Bayesian theories of choice typically

rely on the principle of expected utility, which states that rational decision-makers will choose actions that maximize expected utility. A series of experiments in neuroeconomics illustrates how neural activity can be modeled in Bayesian terms. Platt and Glimcher ′ s recordings of single neurons in the LIP area of the monkey parietal cortex uncovered neurons that are sensitive to probability and to analogs of utility and expected utility.

Checklists

Bayesianism is built on three basic ideas.

1. Belief comes in degrees.

2. Degrees of belief can be modeled as probabilities, which means that they have to conform to the basic principles of the probability calculus.

3. Learning takes place by updating probabilities according to Bayes ′ s rule.

Bayes ′ s rule is a tool for measuring the posterior probability of a hypothesis, conditional upon some evidence.

1. Bayes ′ s rule is calculated from the *prior probability* of the hypothesis, *p* (H); the *likelihood* of the evidence, given the hypothesis, *p* (E/H); and the probability of the evidence, *p* (E).

2.

$$\text{According to Bayes's rule, the posterior probability of the hypothesis} = \frac{\text{Likelihood of the evidence} \times \text{Prior probability of the hypothesis}}{\text{Probability of the evidence}}.$$

3. Bayes ′ s rule is a straightforward consequence of the definition of conditional probability.

Perception can be modeled in terms of Bayesian inference.

1. Perceptual systems have the job of selecting between different hypotheses about the layout of the distal environment on the basis of noisy and partial data (the evidence).

2. This process can be modeled as an application of Bayes ′ s rule because ultimately, perceptual systems have to decide which hypothesis has the highest posterior probability, conditional on the evidence coming from the senses.

Binocular rivalry offers a case study in Bayesian approaches to perception.

1. When separate images are presented to each eye, the perception generated by the visual system alternates between the two images, seemingly at random.

2. From a Bayesian perspective, this is an understandable response to the posterior probabilities calculated via Bayes ′ s rule.

3. The posterior probability for the hypothesis corresponding to the left image (conditional upon the evidence) is the same as the posterior probability for the right-image hypothesis, and both are higher than the posterior probability for the hypothesis that the environment contains a composite corresponding to a blend of the two images.

4. Because the posterior probabilities for the left and right hypotheses are equal, the visual system can ′ t decide between them and so simply

alternates.

Expected utility is the key concept for applying Bayesianism to practical decision-making.

1. Utility is a measure of how much a decision-maker values a particular outcome.

2. To calculate the expected utility of an action, you need to assign utilities to its different possible outcomes and then to those utilities together, each discounted by the probability that it will occur.

3. The St. Petersburg game shows that expected utility is not necessarily the same as expected monetary value.

Platt and Glimcher ' s experiments recording single neurons in the parietal cortex of the monkey are an illustration from neuroeconomics of how the brain may apply Bayesian principles.

1. Neuroeconomics is located at the interface between neuroscience, microeconomics, the psychology of reasoning, behavioral finance, and decision theory.

2. Platt and Glimcher studied how individual neurons in the LIP responded while monkeys made saccadic eye movements for fruit juice rewards.

3. Holding all other aspects of the task constant but varying the probability that each response would be rewarded revealed that the firing rates of individual neurons correlate with both the prior and posterior probabilities of reward.

4. Holding all aspects of the task constant but varying the quantity of the reward revealed that the firing rates of individual neurons correlate with the value of the reward (an analog of utility).

5. Using a free-choice version of the cued-saccade task revealed neurons firing according to the estimated value to the animal of a particular response (as calculated according to Herrnstein ' s melioration theory, which differs from, but is related to, expected utility theory).

Further Reading

Good general introductions to Bayesian approaches to cognitive science can be found in Griffiths et al. 2008 , Chater et al. 2010 , and Jacobs and Kruschke 2011 . For an overview of specific applications, see the special issue of *Trends in Cognitive Sciences* from 2006, devoted to " Probabilistic Models of Cognition. " The essays in Chater and Oaksford 2008 cover a wide range of Bayesian models.

Skyrms 1986 is a classic introduction to inductive logic and theories of probability. Hacking 2001 is more user-friendly. Jeffrey 1983 is a very influential presentation of Bayesian decision theory. Kurt 2019 is an accessible introduction to Bayesian statistics. See Kaplan 1996 for a general philosophical defense of Bayesian approaches to belief and Horwich 1982 for a Bayesian approach to the philosophy of science. For a historical introduction to different ways of thinking about utility, see Broome 1991 . I have written about Bayesian decision theory as a theory of rationality in Bermúdez 2009 .

For Bayesian approaches to perception in general, see the papers in Knill and Whitman 2008 , and for an article-length review focused on visual perception, see Kersten et al. 2004 . Hohwy 2013 uses visual perception and, in

particular, the example of binocular rivalry discussed in Section 7.2 as the starting point for a general theory of the mind as a Bayesian predictive machine. The Bayesian model of binocular rivalry is presented in more detail in Hohwy et al. 2008 . For the neuroscience of binocular rivalry, see de Jong et al. 2020. Clark 2016 discusses predictive coding, relating it to embodied and situated cognition. For further good illustrations of Bayesian approaches, see Wolpert ′ s work on motor control (Körding and Wolpert 2004 , 2006) and Enrst and Banks 2002 on multisensory integration. For a more skeptical view, see Colombo et al. 2021 .

There is an overview of neuroeconomics from a philosophical perspective in Hardy-Vallé 2007 and articles on a range of applications in a special issue of *Brain Research Bulletin* from November 2005. The principal textbook for the field is Glimcher and Fehr 2014 (2nd ed.). Mountcastle et al. 1975 and Robinson et al. 1978 are important early papers in the debate about the specific role of neurons in the posterior parietal cortex. Chapter 10 of Glimcher 2003 accessibly tells the story of single-neuron studies of the LIP, leading up to his own experiments, which are presented more rigorously in Platt and Glimcher 1999 . Kable and Glimcher 2009 reviews the neurophysiology of decisions in primates, whereas Daw et al. 2006 uses functional magnetic resonance imaging to study the neural correlates of exploratory decisions in humans, and Bromberg-Martin and Monosov 2020 reviews the neuroscience of information seeking in uncertain environments. For melioration and rational choice, see Tunney and Shanks 2002 .

Chapter 8

Modules and Architectures



Overview

8.1 Architectures for Artificial Agents

Three Agent Architectures

8.2 Fodor on the Modularity of Mind

Modular and Nonmodular Processing

8.3 The Massive Modularity Hypothesis

The Cheater-Detection Module

The Evolution of Cooperation

Two Arguments

Evaluating the Arguments for Massive Modularity

8.4 Hybrid Architectures: The Example of ACT-R

The ACT-R Architecture

ACT-R as a Hybrid Architecture

Overview

This chapter tackles the overall organization of the mind. Section 8.1 looks at *agent architectures* in artificial intelligence (AI). These are blueprints for the design of artificial agents, which can be anything from robots to internet bots. Looking at different architectures allows us to distinguish cognitive systems from, for example, reflex systems. Reflex systems are governed by simple production rules that uniquely determine how the system will behave in a given situation. In contrast, cognitive systems deploy information processing between the input (sensory) systems and the output (effector) systems.

Cognitive scientists tend to think of the mind (at least in part) as an organized collection of modules, specialized subsystems carrying out specific information-processing tasks. This idea was systematically developed for the first time in *The Modularity of Mind* , written by philosopher Jerry Fodor. We look at Fodor ' s modularity thesis in Section 8.2 .

Fodor distinguishes modular processing from central processing, responsible for general problem solving and decision-making. Massive modularity theorists, in contrast, deny that there is any such thing as nonmodular central processing. We look at this model in Section 8.3 and see how it has been used to explain research into the psychology of reasoning.

Finally, in Section 8.4 , we relate the discussion of agent architectures to different models of information and information processing. We look at an example of a hybrid architecture combining the physical symbol system hypothesis and the parallel-processing characteristic of connectionist

networks. This is the ACT-R architecture, developed by John R. Anderson and colleagues at Carnegie Mellon University.

8.1 Architectures for Artificial Agents

One aim of AI researchers is to build artificial agents. There are many different types of AI agents. Robots are probably the first things to come to mind. Robotic agents are built to operate in real, physical environments. But many agents are designed to function in virtual environments. Shopping bots are good examples. Some bots are designed to travel around the internet comparing prices for a single item, whereas others trawl through sites such as Amazon finding items that you might be likely to buy (perhaps because they have been bought by customers who bought some items that you bought).

Computer scientists have come up with an interesting range of different *agent architectures* for designing artificial agents. An agent architecture is a blueprint that shows the different components that make up an agent and how those components are organized. This section looks at three different types of agent architectures:

- A simple reflex agent
- A goal-based agent
- A learning agent

Looking at these architectures reveals what is distinctive about intelligent, cognitive agents, as opposed to simpler, noncognitive agents. The agent architectures we will be looking at range from the plainly noncognitive to

the plainly cognitive. As we go through them, we get a better picture of the basic functions that any cognitive system has to perform.

First, we need to know what an agent is. The quick definition is that an agent is a system that perceives its environment through *sensory systems* of some type and acts upon that environment through *effector systems* . The basic challenge for a computer scientist programming an agent (whether a software agent or a robotic agent) is to make sure that what the agent does is a function of what the agent perceives. There need to be links between the agent ′ s sensory systems and its effector systems. What distinguishes different types of agents is the complexity of those links between sensory systems and effector systems.

Three Agent Architectures

The simplest type of agent in agent-based computing is the *reflex agent* . Simple reflex agents have direct links between sensory and effector systems. The outputs of the sensory systems directly determine the inputs to the effector systems. These direct links are achieved by *production rules* . Production rules take this form:

IF *condition C holds* THEN *perform action A* .

It is up to the sensory systems to determine whether or not condition C holds. Once the sensory systems have determined that condition C holds, then the behavior of the simple reflex agent is fixed. Figure 8.1 shows a schematic representation of the architecture of a simple reflex agent.

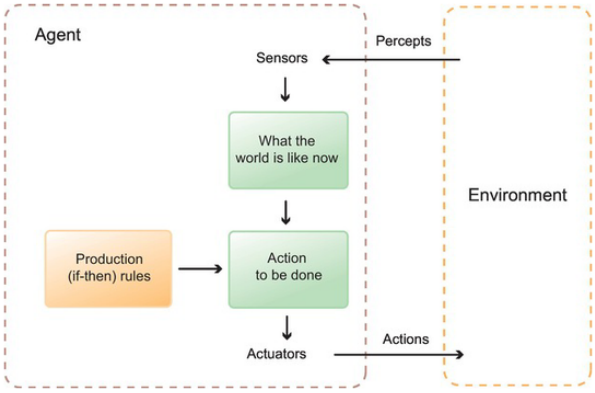


Figure 8.1 The architecture of a simple reflex agent. Production rules are all that intervene between sensory input and motor output.

(Adapted from Russell and Norvig 2009 .)

Simple reflex agents are not cognitive systems because they are simply reacting to the environment in invariant ways. The same stimulus always receives the same response, whereas cognitive systems can react differently to similar environmental stimuli. This is because the actions of cognitive systems are determined by their goals and by their stored representations of the environment. Human agents, for example, sometimes act in a purely reflex manner. But more often, we act as a function of our beliefs and desires — not to mention our hopes, fears, dislikes, and so on.

The schematic agent architecture in Figure 8.2 depicts a *goal-based agent* , which is a primitive type of cognitive system. Goal-based agents do not simply act upon environmental stimuli. There are no simple production rules that will uniquely determine how the agent will behave in a given situation. Instead, such agents need to work out the consequences of different possible actions and then evaluate those consequences in the light of their goals. This is done by the specialized cognitive systems labeled in Figure 8.2 .

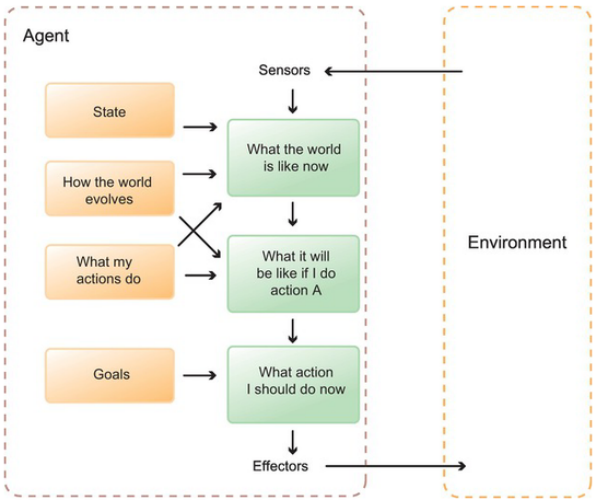


Figure 8.2 The architecture of a goal-based agent. There are information-processing systems intervening between input and output.

(Adapted from Russell and Norvig 2009 .)

But there is still something missing from goal-based agents. They have no capacity to learn from experience, which is surely necessary for any intelligent agent. For that, we need a *learning agent* . A sample architecture for a *learning agent* is presented in Figure 8.3 .

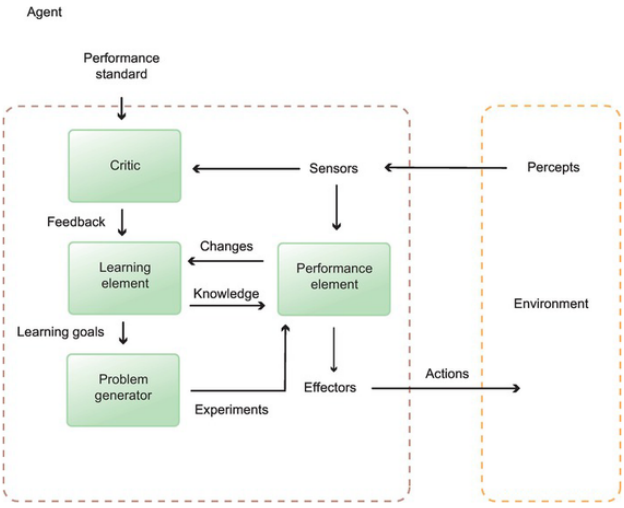


Figure 8.3 The architecture of a learning agent. Mismatches between sensory feedback and the performance standards are detected by the critic subsystem. The learning subsystem determines learning goals and allows the system to experiment with different ways of achieving its goals.

(Adapted from Russell and Norvig 2003 .)

The learning agent has certain standards that it wants its actions to meet. These are one of the inputs to the critic subsystem, which also receives inputs from the sensory systems. The critic ' s job is to detect mismatches between sensory feedback and the performance standard. These mismatches feed into the learning subsystem, which determines learning goals and makes it possible for the system to experiment with different ways of achieving its goals.

As the learning agent example shows, computer scientists designing intelligent agents typically build those agents up from subsystems performing specific information-processing tasks. This way of thinking about cognitive systems (as organized complexes of subsystems) has proven to be very influential in cognitive science. It raises a number of important questions:

- How are we to identify and distinguish cognitive subsystems?
- Are there any important differences between the subsystems responsible for sensory processing and motor behavior, on the one hand, and those that operate between those input and output subsystems?
- Do all the subsystems in a cognitive system process information in the same way? Do they all involve the same type of representations?
- How " autonomous " are the different subsystems? How " insulated " are they from each other?

To explore these questions further, we now turn to the hypothesis that the mind is modular, proposed by the influential philosopher and cognitive scientist Jerry Fodor in his book *The Modularity of Mind* , published in 1983 .

8.2 Fodor on the Modularity of Mind

Modular and Nonmodular Processing

Fodor suggests that there are two fundamentally different types of cognitive processing. *Nonmodular* (or *central*) processes are high level and open-ended; they involve bringing a wide range of information to bear on very general, high-level problems. In contrast, modular processes operate at a much lower level and work quickly to provide rapid solutions to highly determinate problems.

As Fodor defines them, modular processes have the following four characteristics:

- *Domain specificity* . Modules are highly specialized mechanisms designed to carry out specific and circumscribed information-processing tasks. They operate on a limited range of inputs (those relevant to their particular domain).
- *Informational encapsulation* . Modular processing is not affected by what is going on elsewhere in the mind. Modules cannot be " infiltrated " by background knowledge and expectations or by information in the databases associated with different modules.
- *Mandatory application* . Cognitive modules respond automatically to stimuli. They are not under any executive control and cannot be " switched off. " The fact that we cannot avoid perceiving visual illusions even when we know them to be illusions is evidence that certain types of visual processing are modular.
- *Speed* . Modular processing transforms input (e.g., patterns of intensity values picked up by photoreceptors in the retina) into output (e.g., representations of three-dimensional objects) quickly and efficiently.

In addition to these " canonical " characteristics of modular processes, Fodor draws attention to two further features that sometimes characterize modular processes:

- *Fixed neural architecture* . It is sometimes possible to identify determinate regions of the brain associated with particular types of modular processing. So, for example, an area in the fusiform gyrus (the so-called *fusiform face area*) is believed to be specialized for face recognition, which is often described as a modular process.

- *Specific breakdown patterns* . Modular processing can fail in highly determinate ways. These breakdowns can provide clues as to the form and structure of that processing. Prosopagnosia is a highly specific neuropsychological disorder that affects face-recognition abilities but not object recognition more generally.

These last two features are less central because cognitive scientists tend to think of cognitive modules in terms of their function (the information-processing task that they carry out) rather than their physiology. A cognitive module has to perform a single, circumscribed, domain-specific task. But it is not necessary that it map onto a particular part of the brain. Some modules do seem to be localizable, but for others, we have (as yet) no evidence either way.

Here are some mechanisms that Fodor thinks are likely candidates for cognitive modules:

- Color perception
- Shape analysis
- Analysis of three-dimensional spatial relations
- Visual guidance of bodily motions
- Grammatical analysis of heard utterances
- Detection of the melodic or rhythmic structure of acoustic arrays
- Recognition of the voices of conspecifics

Some of these candidate modules, such as color perception, are close to the sensory periphery, with relatively little information processing occurring between the sense organs and the module. Other systems are much further “

downstream. ″ An example here would be grammatical analysis. A lot of processing needs to take place before there is an auditory or visual image to be analyzed for grammatical structure. Moreover, some cognitive modules can take the outputs of other modules as inputs. It is likely that information about the rhythmic structure of an acoustic array will be relevant to identifying the voice of a conspecific, for example.

It seems likely that many motor tasks are also modular. Planning even the simplest reaching movement involves calibrating information about a target object (a glass, say) with information about hand position and body orientation. The location of the glass needs to be coded on a hand-centered coordinate system (as opposed to one centered on the eyes, for example). Executing the movement requires, first, calculating a trajectory that leads from the start location to the end location, then calculating an appropriate combination of muscle forces and joint angles that will take the arm along the required trajectory. These are all highly specialized tasks that seem not to depend on background information or central processing — prime candidates for modular processing, in Fodor ′ s analysis.

But Fodor insists that not all cognition can be carried out by modular mechanisms. There have to be psychological processes cutting across cognitive domains. The very features of cognitive modules that make them computationally powerful, such as their speed and informational encapsulation, mean that their outputs are not always a good guide to the layout of the perceived environment. Appearances can be deceptive. This means that there has to be information processing that can evaluate and correct the outputs of cognitive modules. This is what Fodor and others often call *central processing* to distinguish it from modular processing, which is peripheral.

For Fodor, central processing has two distinguishing features. First, it is *Quinean* (named after the philosopher Willard von Orman Quine, who famously proposed a holistic view of knowledge and confirmation). Each organism ′ s belief system is analogous to a scientific theory. It is, in fact, the organism ′ s theory of the world, and as with scientific theories, it is the belief system as a whole that is evaluated for consistency and coherence, for example. We cannot consider how accurate or well confirmed individual beliefs are in isolation because how we evaluate individual beliefs cannot be divorced from how we think about other elements of the system in which they are embedded.

Second, central processing is *isotropic* . This is a corollary of its Quinean property. To say that central processing is isotropic is, in essence, to say that it is *not* informationally encapsulated. In principle, any part of the belief system is relevant to confirming (or disconfirming) any other. We cannot draw boundaries within the belief system and hope to contain the process of (dis)confirmation within those boundaries.

Fodor himself is very pessimistic about cognitive science being able to shed any light on central processing. He thinks it is best suited to understanding modular processes. It can tell us very little about central processes — about all the processing that takes place in between sensory systems and motor systems. Unsurprisingly, this is not a view that has gained much currency within cognitive science as a whole.

Nonetheless, there are influential movements within cognitive science that are skeptical of the importance and even the existence of what Fodor calls central processing. We have already seen some examples of this from dynamical systems theory in Chapter 6 . The next section looks at a very different approach — the massive modularity hypothesis.

8.3 The Massive Modularity Hypothesis

According to the massive modularity hypothesis, there is no central processing because all information processing is essentially modular. The human mind is a collection of specialized modules, each of which evolved to solve a very specific set of problems that were confronted by our early ancestors — by hunter-gatherers in the Pleistocene period, or even earlier in the evolutionary history of the human ape. These are called *Darwinian modules* .

What sort of Darwinian modules might there be? Evolutionary psychologists have tended to focus primarily on modules solving problems of social coordination, such as problems of cheater detection, kin detection, and mate selection. But massive modularity theorists are also able to appeal to evidence from many different areas of cognitive science pointing to the existence of specialized cognitive systems for a range of different abilities and functions. These include the following:

- Face recognition
- Emotion detection
- Gaze following
- Folk psychology
- Intuitive mechanics (folk physics)
- Folk biology

There is some overlap between Fodorean modules and Darwinian modules, but Darwinian modules typically engage in more complex types of information processing than Fodorean ones. As a consequence, they are

typically not informationally encapsulated. Massive modularity theorists describe Darwinian modules as modular primarily because they are domain-specific modules.

Many different types of evidence are potentially relevant to identifying Darwinian modules. In Chapter 11 , we will look at some influential experiments on prelinguistic infants using the dishabituation paradigm. These experiments show that infants are perceptually sensitive to a number of basic principles governing the behavior of physical objects — such as the principle that objects follow a single continuous path through space and time. These experiments have been taken to show that infants possess a basic theory of the physical world. This basic theory is held by many to be the core of adult folk physics, which itself is the domain of a specialized cognitive system.

The case for massive modularity rests on a mixture of case studies and general arguments. The most developed case study is the so-called *cheater-detection module* . We will look at this first, as a detailed illustration of a Darwinian module. Then we will turn to the general arguments.

The Cheater-Detection Module

We need to start from well-known experiments on reasoning with conditionals (sentences that have an IF . . . THEN . . . structure). These experiments, often using variants of a famous experiment known as the *Wason selection task* , have been widely taken to show that humans are basically very poor at elementary logical reasoning. It turns out, however, that performance on these tasks improves drastically when they are reinterpreted to involve a particular type of conditional. These are so-called

deontic conditionals . Deontic conditionals have to do with permissions, requests, entitlements, and so on. An example of a deontic conditional would be as follows: If you are drinking beer, then you must be over twenty-one years of age.

Evolutionary psychologists Leda Cosmides and John Tooby came up with a striking and imaginative explanation for the fact that humans tend to be much better at reasoning with deontic conditionals than they are with ordinary, nondeontic conditionals. According to Cosmides and Tooby, when people solve problems with deontic conditionals, they are using a specialized module for monitoring social exchanges and detecting cheaters. This is the *cheater-detection module* .

Here is a typical version of the Wason selection task. Imagine that you are shown the four cards illustrated in Figure 8.4 and told that each card has a letter on one side and a number on the other. The experimenter then asks you which cards you would need to turn over in order to determine whether the following conditional is true or false: **If a card has a vowel on one side, then it has an even number on the other** .



Figure 8.4 A version of the Wason selection task. Subjects are asked which cards they would have to turn over in order to determine whether the following conditional is true or false: **If a card has a vowel on one side, then it has an even number on the other** .

It is obvious that the *E* card will have to be turned over. Because the card has a vowel on one side, the conditional will certainly be false if it has an odd number on the other side. Most subjects get this correct. It is fairly

clear that the second card does not need to be turned over, and relatively few subjects think that it does need to be turned over. The problems arise with the two numbered cards.

Reflection shows (or should show!) that the *4* card does not need to be turned over because the conditional would not be disconfirmed by finding a consonant on the other side. The conditional says that any card with a vowel on one side has to have an even number on the other side. It doesn ′ t say anything about cards that have consonants on one side, so the existence of a card with a consonant on one side and an even number on the other is irrelevant.

The *5* card does need to be turned over, though, because the conditional would be falsified if there were a vowel on the other side (this would be a situation in which we have a card with a vowel on one side but no even number on the other).

Unfortunately, very few people see that the *5* card needs to be turned over, whereas the vast majority of experimental subjects think that the *4* card needs to be turned over. This result is pretty robust, as you will find out if you try it on friends and family.

However, it turns out that performance on the selection task varies drastically according to how the task is formulated. There are " real-world " ways of framing the selection task on which the degree of error is drastically diminished. One striking set of results emerged from a variant of the selection task carried out by Richard Griggs and James Cox (1982). They transformed the selection task from what many would describe as a formal test of conditional reasoning to a problem-solving task of a sort familiar to most of the experimental subjects.

Griggs and Cox preserved the structure of the selection task, asking subjects which cards would have to be turned over in order to verify a conditional. But they used a conditional about drinking age rather than one about vowels and even numbers. Subjects were asked to evaluate the following conditional: **If a person is drinking beer, then that person must be over nineteen years of age** (which was, apparently, the law at the time in Florida). They were presented with the cards shown in Figure 8.5 and told that the cards show the names of drinks on one side and ages on the other. Before making their choice, subjects were told to imagine that they were police officers checking whether any illegal drinking was going on in a bar.



Figure 8.5 A version of Griggs and Cox ′ s deontic selection task. Subjects are asked to imagine that they are police officers checking for underage drinkers and are asked which cards they would need to turn over in order to assess the following conditional: **If a person is drinking beer, then that person must be over nineteen years of age** .

The correct answers, of course, are *BEER* and *16* . On this version of the selection task, subjects overwhelmingly came up with the correct answers, and relatively few suggested that the third card would need to be turned over. But it turned out that if the story about the police officers was omitted, performance reverted to a level comparable to that on the original selection task.

The finding that performance on the selection task can be improved by framing the task in such a way that what is being checked is a condition that has to do with permissions, entitlements, and/or prohibitions has proved very

robust. The fact that we are better at reasoning with these *deontic* conditionals than we are with ordinary conditionals has suggested to many theorists that we have a *domain-specific* competence for reasoning involving permissions and prohibitions.

Building on these results, Cosmides and Tooby have suggested that the human mind (perhaps in common with the minds of other higher apes) has a dedicated cognitive system (a *module*) to detect cheaters. They suggest that when the selection task is framed in terms of permissions and entitlements, it engages the cheater-detection module. This is why performance suddenly improves .

The Evolution of Cooperation

But why should there be a cheater-detection module? What was the pressing evolutionary need to which the cheater-detection module was a response? Cosmides and Tooby answer these questions through an influential theory of the emergence of cooperative behavior.

Biologists, and evolutionary theorists more generally, have long been puzzled by the problem of how cooperative behavior might have emerged from a process of natural selection. Cooperative behavior presumably has a genetic basis. But how could the genes that code for cooperative behavior ever have become established if an individual who takes advantage of cooperators without reciprocating will always do better than one who cooperates? Evolution seems to favor *free riders* and exploiters above high-minded altruists.

A popular way of thinking about the evolution of cooperation is through the model of the *prisoner ' s dilemma* (explained in Box 8.1). Many

interpersonal interactions (and for that matter, many interactions between nonhuman animals) involve a series of encounters where each has the structure of a prisoner ′ s dilemma, but it is not known how many encounters there will be. Game theorists call these *indefinitely iterated* prisoner ′ s dilemmas.

BOX 8.1 The Prisoner ′ s Dilemma

A prisoner ′ s dilemma is a strategic interaction in which, if each participant does what seems to be the rational thing from their individual perspective, then the result for everyone is much worse than they could have achieved by cooperating.

A police officer is interrogating two prisoners. He proposes to each of them that they betray the other and explains the possible consequences. If each prisoner betrays the other, then they will both end up with a sentence of five years in prison. If neither betrays the other, then they will each be convicted of a lesser offense and both end up with a sentence of two years in prison. If either prisoner betrays the other without himself being betrayed, however, then he will go free while the other receives ten years in prison. Here is the payoff table:

		PLAYER B	
		Betray	Not betray
PLAYER A	Betray	–5, –5	0, –10
	Not betray	–10, 0	–2, –2

The outcomes are given in terms of the number of years in prison that will ensue for prisoners A and B, respectively (presented as a negative number because years in prison are undesirable). So, the outcome in the bottom left-hand box is ten years in prison for prisoner A and none for prisoner B, which occurs when prisoner A does not betray but prisoner B does.

Prisoner A might reason like this: *Prisoner B can either betray me or not. Suppose he betrays me. Then I have a choice between five years in prison if I also betray him — or ten years if I keep quiet. So, my best strategy if he betrays me is to betray him. But what if he does not betray? Then I have a choice between two years if I keep quiet as well or going free if I betray him. So, my best strategy if he is silent is to betray him. Whatever he does, therefore, I ′ m better off betraying him.*

A game theorist would say that betrayal is prisoner A ′ s *dominant strategy* . It promises a greater advantage to that individual than the other available strategies, irrespective of what the other player does.

Unfortunately, things look exactly the same to prisoner B. So, the two prisoners will end up betraying each other and spending five years each in prison, even though they could have been better off keeping silent and spending two years each in prison .

One way of dealing with repeated social interactions of this kind is to adopt a simple heuristic strategy in which one bases one ′ s plays not on how one expects others to behave but, rather, on how they have behaved in the past. The best known of these heuristic strategies is TIT FOR TAT, which is composed of the following two rules:

1. Always cooperate in the first encounter.

2. In any subsequent encounter, do what your opponent did in the previous round.

Theorists have found TIT FOR TAT a potentially powerful explanatory tool in explaining the evolutionary emergence of altruistic behavior for two reasons. First, it is simple and involves no complicated calculations. And second, it is what evolutionary game theorists call an *evolutionarily stable strategy* — that is to say, a population where there are sufficiently many " players " following the TIT FOR TAT strategy with a sufficiently high probability of encountering each other regularly will not be invaded by a subpopulation playing another strategy (such as the strategy of always defecting). TIT FOR TAT, therefore, combines simplicity with robustness.

Here, finally, we get to the cheater-detection module. Applying TIT FOR TAT requires being able to identify instances of cooperation and defection. It involves being able to tell when an agent has taken a benefit without paying the corresponding price. An agent who consistently misidentifies defectors and free riders as cooperators (or, for that matter, vice versa) will not flourish.

This is why, according to Cosmides and Tooby, we evolved a specialized module in order to allow us to navigate social situations that depend crucially on the ability to identify defectors and free riders. Because the detection of cheaters and free riders is essentially a matter of identifying when a conditional obligation has been breached, this explains why we are so much better at deontic versions of the selection task than ordinary versions — and why we are better, more generally, at conditional reasoning about rules, obligations, and entitlements than we are at abstract conditional reasoning .

Exercise 8.1 Explain the argument from the evolution of cooperation to the cheater-detection module in your own words.

Two Arguments

The extended case study of the cheater-detection module is reinforced by two more general arguments, to which we turn now.

Starting from the facts that the human mind is the product of evolution and that evolution works by natural selection, the two arguments try to show that evolution could not have selected a domain-general mental architecture. No domain-general, central processing system of the type that Fodor envisages could have been selected because no such processing system could have solved the type of adaptive problems that fixed the evolution of the human mind.

The argument from error. It is a basic fact that natural selection selects for heritable traits that preserve fitness. But what counts as fitness? What are the criteria for fitness?

According to Cosmides and Tooby, these fitness criteria have to be domain-specific, not domain-general, criteria. What counts as fitness-promoting behavior varies from domain to domain. They give the example of how one treats one ′ s family members. It is certainly not fitness-promoting to have sex with close family members. And it is fitness-promoting to help family members in many other circumstances. But not always. If one is in a social exchange with a prisoner ′ s dilemma – type structure and is applying something like the TIT FOR TAT algorithm, then it

is only fitness-promoting to help family members who are cooperating — not the ones who are taking the benefit without paying the costs.

So, because there are no domain-general fitness criteria, there cannot (they argue) be domain-general cognitive mechanisms. Domain-general cognitive mechanisms could not have been selected by natural selection because they would have made too many mistakes — whatever criteria of success and failure they had built into them would have worked in some cases but failed in many more. Instead, say Cosmides and Tooby, there must be a distinct cognitive mechanism for every domain that has a different definition of what counts as a successful outcome.

Exercise 8.2 State the argument from error in your own words and evaluate it.

The argument from statistics and learning. Like the previous argument, this argument focuses on problems in how domain-general cognitive systems can discover what fitness consists of. The principal problem is that the world has what Cosmides and Tooby describe as a "statistically recurrent domain-specific structure." Certain features hold with great regularity in some domains but not in others. These are not the sorts of things that a general-purpose cognitive mechanism could be expected to learn.

Their example is the model of kin selection proposed by evolutionary biologist W. D. Hamilton (Figure 8.6). The problem of kin selection is the problem of explaining why certain organisms often pursue strategies that promote the reproductive success of their relatives at the cost of their own

reproductive success. This type of self-sacrificing behavior seems, at first glance, to fly in the face of the theory of natural selection because the self-sacrificing strategy seems to diminish the organism ' s fitness.



Figure 8.6 The evolutionary biologist W. D. Hamilton (1936 – 2000).

(Photo by Jeffrey Joy.)

Hamilton ' s basic idea is that there are certain circumstances in which it can make good fitness-promoting sense for individuals to sacrifice themselves for another individual. From an evolutionary point of view, fitness-promoting actions are ones that promote the spread of the agent ' s genes. And, Hamilton argued, there are circumstances where an act of self-sacrifice will help the individual ' s own genes to spread and thereby spread the kin-selection gene. In particular, this holds when the following are true:

Condition 1: The self-sacrificer shares a reasonable proportion of genes with the individual benefiting from the sacrifice.

Condition 2: The individual benefiting from the sacrifice shares the gene that promotes kin selection.

What counts as a reasonable proportion? According to Hamilton ′ s famous kin-selection equation, kin-selection genes will spread when the following holds:

$$R_{xy}B_y > C_x.$$

Here, the x subscript refers to the self-sacrificer, and the y subscript refers to the beneficiary of the sacrifice. R_{xy} measures how related x and y are. C_x measures the reproductive cost of kin selection to x , and B_y measures the reproductive benefit to y . In English, therefore, Hamilton ′ s kin-selection equation says that kin-selection genes will spread when the reproductive benefit to the recipient of the sacrifice, discounted by the recipient ′ s degree of relatedness to the self-sacrificer, exceeds the reproductive cost to the self-sacrificer.

Typically, two sisters will share 50 percent of their genes — or, more precisely, 50 percent of the variance in their genes (i.e., what remains after taking away all the genetic material likely to be shared by any two randomly chosen conspecifics). So, if x and y are sisters, then $R_{xy} = 0.5$. This tells us that it is only fitness-promoting for one sister to sacrifice her reproductive possibilities to help her sister when her sister will thereby do twice as well (reproductively speaking!) as she herself would have done if she hadn ′ t sacrificed herself. So, the sacrifice will be fitness-promoting if, for example,

the self-sacrificing sister could only have one more child, but the sacrifice enables her sister to have three more.

Cosmides and Tooby think that massive modularity is the only way of explaining how the kin-selection law got embedded in the population. The kin-selection equation exploits statistical relationships that completely outstrip the experience of any individual. No domain-general learning mechanism could ever pick up on the statistical generalizations that underwrite Hamilton ' s kin-selection law.

The only way the kin-selection law could have become embedded in the population would be if natural selection had selected a special-purpose kin-selection module with the law built into it.

Exercise 8.3 State the argument from statistics and learning in your own words and evaluate it .

Evaluating the Arguments for Massive Modularity

Both the argument from error and the argument from statistics and learning are compatible with the idea that human beings (not to mention other animals) are born with certain innate bodies of domain-specific *knowledge* . This is important because information processing can exploit domain-specific knowledge without being modular.

When we are thinking about the organization of the mind, the distinction between domain-specific modules and domain-specific knowledge is very important. Formulated in terms of cognitive modules, the massive modularity hypothesis is a bold and provocative doctrine about the

overall structure of the mind. It says that there is no such thing as a domain-general information-processing mechanism and that the mind is nothing over and above a collection of independent and quasi-autonomous cognitive subsystems.

But formulated in terms of domain-specific bodies of knowledge, it is much less controversial. The idea that we (and quite possibly other animals) are born with innate bodies of knowledge dedicated to certain domains is not really a claim about the architecture of cognition. Cognitive scientists have proposed such innate bodies of knowledge in a number of different areas — such as numerical competence, intuitive mechanics, and so on.

Nonetheless, even if one does not accept the massive modularity hypothesis in its strongest form, it still makes some very important points about the organization of the mind. In particular, it makes a case for thinking that the mind might be at least partially organized in terms of cognitive subsystems or modules that are domain-specific modules without having all the characteristics of full-fledged Fodorean modules. Cognitive scientists have taken this idea very seriously, and we will be exploring it further in later chapters.

Chapter 13 , for example, develops a case study that brings the theoretical discussions about modularity to life. We will look at a debate that is very much at the forefront of contemporary cognitive science — the controversial question of whether there is a module responsible for reasoning about the mental states of others, or what many cognitive scientists have come to call the *theory-of-mind module* .

8.4 Hybrid Architectures: The Example of ACT-R

This discussion of modules connects with earlier discussions of information processing. The distinction between physical symbol systems and artificial neural networks is not an all-or- nothing distinction. Symbolic and distributed information processing seem to be suited for different tasks and for solving different types of problems.

The types of problems tackled by good old-fashioned artificial intelligence (GOFAI) physical symbol systems tend to be highly structured and sharply defined — playing checkers, for example, or constructing decision trees from databases. The types of problems for which artificial neural networks seem particularly well suited tend to be perceptual (distinguishing mines from rocks, for example, or modeling how infants represent unseen objects) and involve recognizing patterns (such as patterns in forming the past tense of English verbs).

The extreme version of the physical symbol system hypothesis holds that *all* information processing involves manipulating and transforming physical symbol structures. It may be that Newell and Simon themselves had something like this in mind. There is a comparable version of the artificial neural networks approach, holding that physical symbol structures are completely redundant in modeling cognition — artificial neural networks are all we need. There seems to be room, though, for a more balanced approach that tries to incorporate both models of information processing. The ACT-R cognitive architecture developed by Michael Anderson and his research team at Carnegie Mellon University is a good example of how this might work.

A cognitive architecture is similar to a programming language. It gives researchers the tools to construct cognitive models using a common language and common toolkit. One of the first cognitive architectures was actually developed by Allen Newell, working with John Laird and Paul Rosenbloom. It was originally called *SOAR* (for " state operator and result "). The current incarnation is known as *Soar* . Soar is very closely tied to the physical symbol system hypothesis. It is based on the means – end and heuristic search approaches to problem solving that we looked at in Chapter 4 . Soar is intended to be a unified model of cognition. It does not incorporate any elements corresponding to artificial neural networks. All knowledge is represented in the same way in the architecture and is manipulated in a rule-governed way.

But why not think about the mind in terms of *both* ways of modeling information processing, with some modules engaged in symbolic information processing and others engaged in distributed information processing? That is the basic idea behind the ACT-R cognitive architecture developed by psychologist Anderson and his research team at Carnegie Mellon.

The ACT-R cognitive architecture is the latest installment of a cognitive architecture that was first announced under the name *ACT* in 1976 (*ACT* stands for " adaptive control of thought " and *R* for " rational "). It is a hybrid architecture because it incorporates both symbolic and subsymbolic information processing. The ACT-R architecture has a modular organization. Different modules perform different cognitive tasks, and the type of information processing in a module depends on the type of task the module performs.

The ACT-R Architecture

The basic structure of ACT-R is illustrated in Figure 8.7 . I want to draw your attention, in particular, to the two boxes at the top (the perceptual module and the motor module) and to the two modules directly below them (procedural memory and declarative memory).

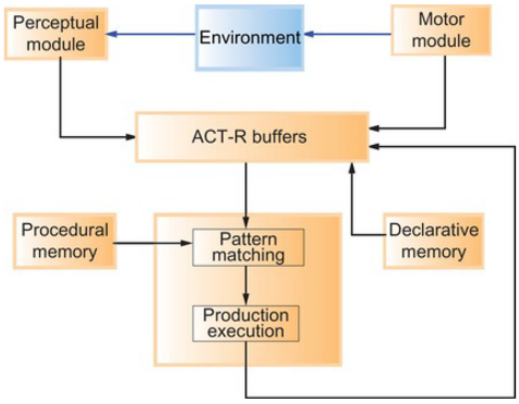


Figure 8.7 The ACT-R cognitive architecture.

In some versions of ACT-R, the perceptual module is actually composed of a range of distinct modules — the visual module, the auditory module, and so on, depending on the number of different types of input that the architecture can receive. Likewise, the motor module can sometimes itself be broken down into a speech module, a manual module, and so on, depending on the different types of action that the architecture can perform. It is probably more helpful to talk generally about the *perceptual – motor layer* . Declarative and procedural memory collectively constitute what we can term the *cognitive layer* .

The cognitive layer incorporates two fundamentally different types of knowledge — declarative and procedural. In philosophy, this is often labeled as the distinction between *knowledge-that* (declarative) and *knowledge-how*

(procedural) — between, for example, knowing that Paris is the capital of France and knowing how to speak French.

Exercise 8.4 Explain the distinction between knowledge-that and knowledge-how in your own words.

Declarative and procedural knowledge are both represented symbolically but in different ways. Declarative knowledge is organized in terms of " chunks. " A chunk is an organized set of elements. These elements may be derived from the perceptual systems, but chunks can themselves be chunked. We can think of chunks as symbol structures (say, the equation **7 + 6 = 13**) built up in rule-governed ways from physical symbols (corresponding to **7** , **6** , + , **1** , and **3**). These chunks are stored in the declarative memory module.

ACT-R uses *production rules* to encode procedural knowledge. Production rules identify specific actions for the system to perform, depending on which condition it finds itself in. When a production rule fires (as the jargon has it) in a given condition, it can perform one of a range of actions. It can retrieve a chunk from declarative memory, for example. Or it can modify that chunk — updating its representation of the environment, for example, or modifying a goal. It can also send a command to the motor module. Production rules can be nested within each other so that the output of a given production rule serves as a condition triggering the firing of another production rule. This allows complex abilities (such as multiplication) to be modeled as sets of production rules.

So far, there is nothing hybrid about ACT-R. The way declarative and procedural knowledge is encoded and manipulated in the architecture is entirely in line with the physical symbol system hypothesis. And in fact, the same holds for the perceptual and motor modules. Here, too, information is encoded in the form of physical symbols. In some versions of ACT-R, the perceptual and motor modules are designed on the basis of the Executive Process/Interactive Control (EPIC) architecture developed by David Kieras and David Meyer. EPIC falls squarely within the physical symbol system approach.

ACT-R as a Hybrid Architecture

What makes ACT-R a hybrid architecture is that this symbolic, modular architecture is run on a subsymbolic base. Take another look at Figure 8.7 . In many ways, the overall organization looks very Fodorean. There are various modules, and they are all encapsulated. They communicate only via buffer systems that filter out all nonrelevant inputs. And yet there is something missing. There is no system responsible for what Fodor would call central processing.

So, a natural question to ask of ACT-R is: How does it decide what to do? This is really the question of how it decides which production rules to apply. ACT-R is designed to operate serially. At any given moment, only one production rule can be active. But most of the time, there are many different production rules that *could* be active. Only one of them is selected. How?

The job of selecting which production rule is to be active at a given moment is performed by the *pattern-matching module* . This module controls which production rule gains access to the buffer. It does this by

working out which production rule has the highest utility at the moment of selection.

The production rule with the highest utility is the rule whose activation will best benefit the cognitive system. What counts as a benefit depends on the system ′ s goals — or rather, the system ′ s current goal. The utility of a particular production rule is determined by two things: how likely the system is to achieve its current goal if the production rule is activated and, second, the cost of activating the production rule.

So, the pattern-matching module essentially carries out a form of cost – benefit analysis to determine which production rule should gain access to the buffer. This cost – benefit calculation is really an application of the Bayesian approach discussed in Chapter 7 . In effect, the pattern-matching module is calculating expected utility (and then factoring in the costs of applying the production rule).

The entire process takes place without any overseeing central system. It is a type of " winner-take-all " system. All the work is done by the equations that continually update the cost and utility functions. Once the numbers are in, the outcome is determined.

The designers of ACT-R describe these calculations as *subsymbolic* . This is a very important concept that is also standardly used to describe how artificial neural networks operate. For contrast, note that each production rule is purely symbolic. Production rules are built up in rule-governed ways from basic constituent symbols exactly as the physical symbol system hypothesis requires.

The compositional structure of production rules determines how the production rule behaves once it is activated, but it does not play a part in determining whether or not the rule is activated. For that, we need to turn to

the numbers that represent the production rule ' s utility. These numbers are subsymbolic because they do not reflect the symbolic structure of the production rule.

ACT-R has other subsymbolic dimensions, as summarized in Table 8.1 . For example, it uses subsymbolic equations to model how accessible information is in declarative memory. The basic units of declarative memory are chunks — as opposed to the production rules that are the basic units of procedural memory. Each chunk has associated with it a particular activation level. This activation level can be represented numerically. The higher the activation level, the easier it is to retrieve the chunk from storage.

Table 8.1 Comparing the symbolic and subsymbolic dimensions of knowledge representation in the hybrid ACT-R architecture

	PERFORMANCE MECHANISMS		LEARNING MECHANISMS	
	SYMBOLIC	**SUBSYMBOLIC**	**SYMBOLIC**	**SUBSYMBOLIC**
Declarative chunks	Knowledge (usually facts) that can be directly verbalized	Relative activation of declarative chunks affects retrieval	Adding new declarative chunks to the set	Changing activation of declarative chunks and changing strength of links between chunks
Production rules	Knowledge for taking particular actions in particular situations	Relative utility of production rules affects choice	Adding new production rules to the set	Changing utility of production rules

The activation levels of chunks in declarative memory are determined by equations similar to those governing the utilities of production rules. There are two basic components determining a chunk ' s overall activation level. The first component has to do with how useful the chunk has been in the past. As before, usefulness is understood in terms of utility, which in turn is understood in terms of how the chunk has contributed to realizing the system ' s goals. The second component has to do with how relevant the chunk is to the current situation and context.

The example of ACT-R reveals two important lessons. First, debates about the organization of the mind are closely connected to debates about the nature of information processing. Thinking properly about the modular organization of the mind requires thinking about how the different modules might execute their information-processing tasks. And second, the different parts of a mental architecture might exploit different models of information processing. Some tasks lend themselves to a symbolic approach, others to a subsymbolic approach. The debate between models of information processing is not an all-or-nothing case.

Summary

This chapter has focused on different ways of thinking about how the mind is organized. We began by looking at three different architectures for intelligent agents in AI, in order to see what distinguishes cognitive agents from simple reflex agents. Cognitive agents are standardly modeled in terms of quasi-autonomous information-processing systems, which raises the question of how those systems should be understood. Pursuing this question took us to Jerry Fodor ′ s analysis of modular information-processing systems and explored his reasons for thinking that cognitive science is best suited to explaining modular systems, as opposed to nonmodular, central information-processing systems. We then examined an alternative proposed by massive modularity theorists, who hold that all information processing is modular. Finally, we turned to the hybrid architecture ACT-R, which brings the discussion of modularity into contact with the discussion of information processing in Part III. ACT-R is a modular system that combines the

symbolic approach associated with the physical symbol system hypothesis and the subsymbolic neural networks approach.

Checklists

Computer scientists building intelligent agents distinguish different types of agent architectures.

1. Simple reflex agents have condition – action rules (production rules) directly linking sensory and effector systems.

2. Simple reflex agents are not cognitive systems, unlike goal-based agents and learning agents.

3. Goal-based agents and learning agents are built up from subsystems that perform specific information-processing tasks.

4. This general approach to agent architecture raises theoretical questions explored in discussions of modularity.

Fodor ' s Modularity Thesis

1. The thesis proposes that there are specialized information-processing modules that are domain specific, informationally encapsulated, mandatory, and fast.

2. These modules may also have a fixed neural architecture and specific breakdown patterns.

3. Modules are employed for basic types of information processing (e.g., shape analysis, color perception, and face recognition).

4. Modules provide inputs to nonmodular, central processing — the realm of belief fixation and practical decision-making, among other things.

5. Central processing is holistic and so not informationally encapsulated.

According to the massive modularity hypothesis, all information processing is modular. There is no domain-general information processing.

1. The human mind is claimed to be a collection of specialized modules, each of which evolved to solve a specific set of problems encountered by our Pleistocene ancestors.

2. Examples of these Darwinian modules are the cheater-detection module and modules proposed for folk psychology (theory of mind) and folk physics (intuitive mechanics).

3. According to the argument from error, domain-general cognitive mechanisms could not have evolved because there are no domain-general fitness criteria.

4. According to the argument from statistics and learning, domain-general learning mechanisms cannot detect statistically recurrent domain-specific patterns (such as the kin-selection equation proposed by W. D. Hamilton).

5. However, both of these arguments can be satisfied with the much weaker claim that there are innate, domain-specific bodies of knowledge.

ACT-R is an example of a hybrid architecture that combines both symbolic and subsymbolic elements.

1. In ACT-R, knowledge is represented in two different ways — declarative knowledge is represented in chunks, whereas procedural knowledge is represented through production rules.

2. Items of knowledge become available for general information processing when they appear in one of the buffers. This general information processing is fundamentally symbolic in character.

3. In contrast, the processes that determine whether a particular item of knowledge ends up in a buffer are subsymbolic — equations, for example, that calculate how useful a given production rule might be in a particular context.

4. These processes are subsymbolic because they do not exploit or depend on the internal symbolic structure of the item of knowledge.

Further Reading

There is a useful introduction to intelligent agents in Russell and Norvig 2009 , particularly Chapter 2 . An earlier version of this chapter (from the book ' s first edition) is available in the online resources. A good review can also be found in Poole and Mackworth 2010 . See the online resources for other helpful collections pertaining to agent architectures. Lieto 2021 explores how design principles from cognitive science can be used in developing artificial minds.

Fodor ' s modularity thesis is presented in his short book *The Modularity of Mind* (1983). A summary of the book, together with peer

commentaries, was published in the journal *Behavioral and Brain Sciences* (Fodor 1985). The summary is reprinted in Bermúdez 2006 . For critical discussion of the modularity of face perception, see Kanwisher et al. 1997 and Kanwisher 2000 . Allott and Smith 2021 compares Fodor and Chomsky on modularity. Clarke 2021 defends the general idea that there are informationally encapsulated cognitive systems. On domain specificity, see Spunt and Adolphs 2017 .

Cosmides and Tooby have written an online evolutionary psychology primer, available in the online resources. More recent summaries of Cosmides and Tooby ' s research can be found in Cosmides et al. 2010 and Cosmides and Tooby 2013 . Their 1994 paper discussed in the text (Cosmides and Tooby 1994) is reprinted in Bermúdez 2006 . It was originally published in Hirschfeld and Gelman 1994 . This influential collection contains a number of other papers arguing for a modular approach to cognition. There is a useful entry on biological altruism in the online *Stanford Encyclopedia of Philosophy* (https://plato.stanford.edu/). For Hamilton ' s theory of kin selection, see Dawkins 1979 (available in the online resources), and for a recent application to explaining the evolution of cooperation among gut microbes, see Simonet and McNally 2021 .

Pinker 1997 develops a view of the mind that integrates the massive modularity hypothesis with other areas of cognitive science. Pinker is a particular target of Fodor ' s discussion of massive modularity in Fodor 2000 . Pinker responds to Fodor in Pinker 2005 (available in the online resources). For a critical perspective on the entire debate about massive modularity, see Pietraszewski and Wertz 2021 . Burkart et al. 2017 discusses different approaches to the evolution of general intelligence, whereas Zelditch and Goswami 2021 explores the different conceptions of modularity current in

discussions of evolution (which are rather different from those discussed in this chapter).

Carruthers 2006 is a book-length defense of a version of the massive modularity thesis. The journal *Mind and Language* published a precis of the book (Carruthers 2008b), together with three commentaries — Machery 2008 , Wilson 2008 , and Cowie 2008 . Carruthers replies in the same issue (Carruthers 2008a). A good review of modularity research can be found in Barrett and Kurzban 2006 . Also see Richard Samuels ' s chapter on massive modularity in Margolis et al. 2012 . The *Stanford Encyclopedia of Philosophy* (https://plato.stanford.edu/) also has an entry on modularity.

Kotseruba and Tsotsos 2020 reviews forty years of research on cognitive architectures in general. The home page for the ACT architecture is the best place to start (see online resources). It contains a comprehensive bibliography with links to PDF versions of almost every referenced paper. For a brief overview of the general ACT approach, see Lebiere 2003 . For a longer introduction to ACT-R, see Anderson et al. 2004. To see how ACT-R can be implemented neurally, see Zylberberg et al. 2011 . Dimov et al. 2020 uses ACT-R to develop a neurocognitive model of decision-making.

Chapter 9

Strategies for Brain Mapping



Overview

9.1 Structure and Function in the Brain

Exploring Anatomical Connectivity

9.2 Studying Cognitive Functioning: Techniques from Neuroscience

Mapping the Brain ′s Electrical Activity: EEG and MEG

Mapping the Brain ′s Blood Flow and Blood Oxygen Levels: PET and fMRI

9.3 Combining Resources I: The Locus-of-Selection Problem

Combining ERPs and Single-Unit Recordings

9.4 Combining Resources II: Networks for Attention

Two Hypotheses about Visuospatial Attention

9.5 From Data to Maps: Problems and Pitfalls

From Blood Flow to Cognition?

Noise in the System?

Functional Connectivity versus Effective Connectivity

Overview

This chapter explores the enterprise of making a wiring diagram for the mind. Neuroanatomy (the study of the anatomical structure of the brain) is a good place to start, but it can only take us so far. We are looking for a *cognitive* wiring diagram. This takes us beyond anatomy because cognitive functions rarely map cleanly onto brain areas. Section 9.1 looks in more detail at the theoretical and practical issues that arise when we start to think about the interplay between *structure* and *function* in the brain.

Many neuroscientists think that we can *localize* particular cognitive functions in specific brain areas (or networks of brain areas). Their confidence is in large part due to the existence of powerful techniques for studying patterns of cognitive activity in the brain. These techniques include the following:

- Electroencephalography (EEG) for measuring event-related potentials (ERPs)
- Positron emission tomography (PET)
- Functional magnetic resonance imaging (fMRI)

Section 9.2 introduces these techniques and their respective strengths, while the case studies in Sections 9.3 and 9.4 show how the different techniques can be combined to shed light on the complex phenomenon of attention.

Neuroimaging techniques do not provide a direct "window" on cognitive functions. They provide information about blood flow (in the case of PET) or the blood-oxygen-level–dependent (BOLD) signal (in the case of fMRI). How we get from there to models of cognitive organization

depends on how we interpret the data. Section 9.5 looks at some of the challenges this involves.

9.1 Structure and Function in the Brain

The brain has conspicuous anatomical landmarks. Most obviously, it has two hemispheres, left and right, with the division between them going lengthwise down the middle of the brain. As we saw in Section 3.2 , the cortical surface of each of these hemispheres is divided into four lobes. Each of the four lobes is thought to be responsible for a different type of cognitive functioning. The frontal lobe is generally associated with reasoning, planning, and problem solving, for example. Anatomically speaking, however, the lobes are distinguished by large-scale topographic features known as *gyri* and *sulci* (singular: *gyrus* and *sulcus*).

If you look at a picture of the surface of the brain, you will see many bumps and grooves. The bumps are the gyri, and the grooves are the sulci. The sulci are also known as *fissures* . Many of these bumps and grooves have names. Some names are purely descriptive. The parieto-occipital sulcus, for example, separates the parietal lobe from the occipital lobe. Some of the names are more interesting. The Sylvian sulcus (which is marked in Figure 9.1 as the lateral cerebral sulcus) divides the temporal lobe from the lobe in front of it (the frontal lobe) and from the lobe above it (the parietal lobe). It is named after Franciscus Sylvus, who was a seventeenth-century professor of medicine at the University of Leiden in the Netherlands.

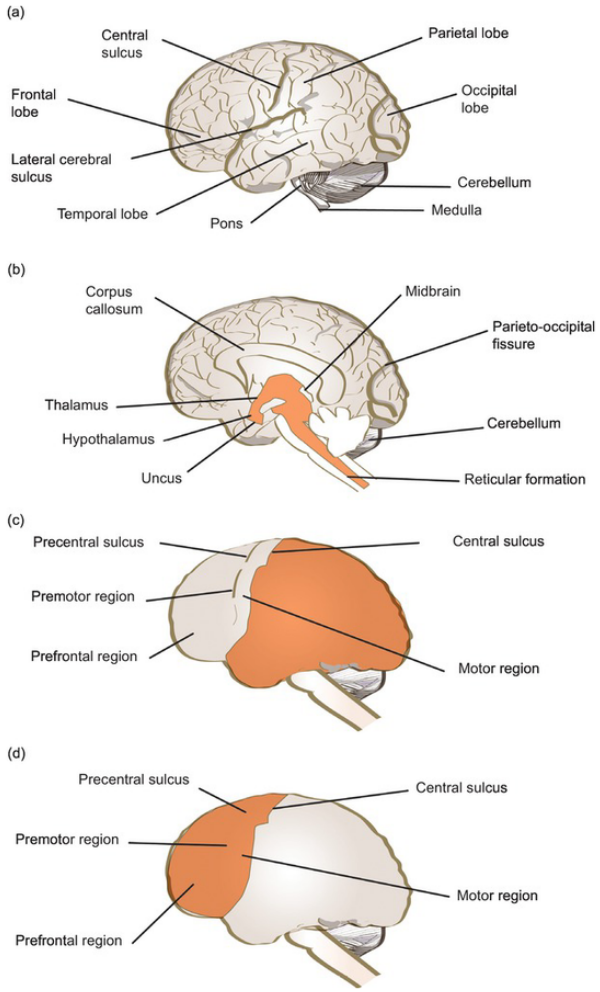


Figure 9.1 Luria ′ s (1970) diagram of the functional organization of the brain. The top diagram is anatomical, whereas the other three depict functional networks.

(Adapted from Luria 1970 .)

The diagram in Figure 9.1a is drawn from a review article published in *Scientific American* in 1970 by the famous Russian neuropsychologist Alexander Luria. It illustrates some of the most prominent large-scale features of the anatomy of the brain ′ s surface. My main interest in reproducing it, however, is to contrast it with the other three diagrams in Figure 9.1 . Each of these depicts one of what Luria thought of as the three

main *functional networks* in the brain. Luria called these networks " blocks. " They are colored brown in the diagrams.

According to Luria, each block has very different roles and responsibilities. Figure 9.1b is the most primitive block, made up of the brain stem and the oldest parts of the cortex. (This would be a good moment to look back at the first few paragraphs of Section 3.2 .) According to Luria, this system regulates how awake and responsive we are. The second block (in Figure 9.1c) regulates how we code, control, and store information, whereas the third block (Figure 9.1d) is responsible for intentions and planning.

The specific details of Luria ' s analysis are not particularly important. What he was reviewing in 1970 is no longer state of the art now. But Luria ' s diagram clearly illustrates two things. The first is the difference between anatomy and cognitive function. The diagram in Figure 9.1a is an anatomical diagram. It divides the brain into regions on the basis of large-scale anatomical features, but it has nothing to say about what those regions actually do.

In contrast, the other three diagrams mark many of the same anatomical regions, but they are organized in functional terms. This is particularly clear in Figures 9.1c and 9.1d , corresponding to Luria ' s second and third blocks. Here, we have regions picked out in terms of what they are thought to do (in terms of the cognitive function that they serve). So, for example, a particular section of the frontal lobe is identified as the motor region (responsible for planning voluntary movements).

The second thing that we learn from Luria ' s diagram is how easy it is to slide from talking about anatomical areas to talking about functional areas (and vice versa). When we talk about the Sylvian sulcus, we are talking

about an anatomical feature of the brain. When we talk about the motor region, in contrast, we are talking about a region of the brain identified in terms of its function. But it is very common to have (as we have here) diagrams and maps of the brain that use both types of labels. And in fact, the same area can have two very different names, depending on how we are thinking about it. The precentral gyrus, for example, is an anatomical feature located just in front of the central sulcus. It is also called the *primary motor cortex* because neuroscientists have discovered that directly stimulating this area causes various parts of the body to move.

Exploring Anatomical Connectivity

The *principle of segregation* is fundamental to neuroscience. This is the idea that the cerebral cortex is divided into segregated areas with distinct neuronal populations. Neuroscientists still use a classification of anatomical areas in the cerebral cortex developed by the great German neuroanatomist Korbinian Brodmann in the late nineteenth and early twentieth centuries.

Brodmann distinguished different regions in the cerebral cortex in terms of the types of cells that they contain and how densely those cells occur. To study cell distribution in the cortex, Brodmann used recently discovered techniques for staining cells. Staining methods are still used by neuroscientists today. They involve dipping very thin slices of brain tissue into solutions that allow details of cellular structure to be seen under a microscope. Brodmann used the Nissl stain, developed by the German neuropathologist Franz Nissl. The Nissl stain turns all cell bodies a bright violet color.

Using the Nissl stain to examine the distribution of different types of neurons across the cerebral cortex, Brodmann identified over fifty different cortical regions. Figure 9.2 gives two views of the brain with the four lobes and the different Brodmann areas clearly marked. The top view is a lateral view (from the side), whereas the lower one is a medial view (down the middle).

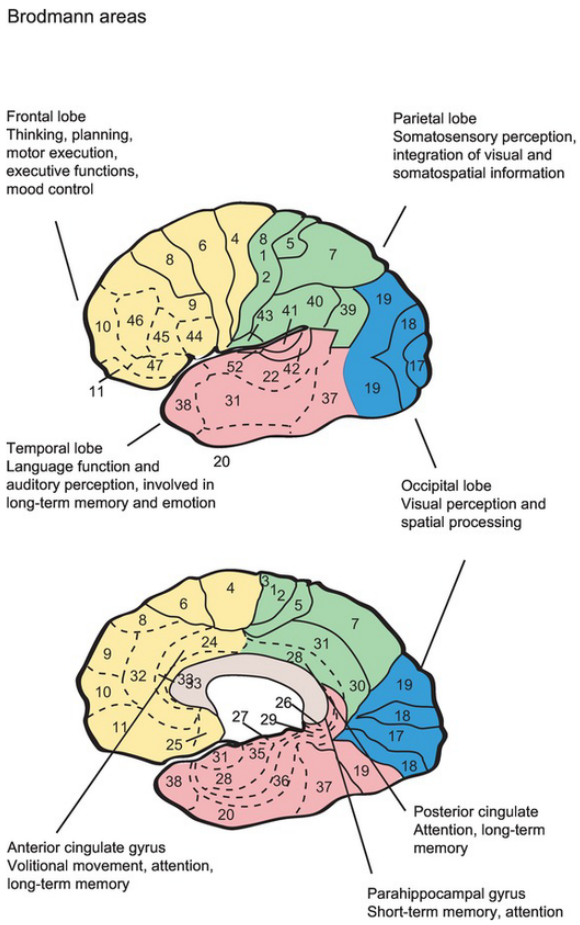


Figure 9.2 Map of the anatomy of the brain showing the four lobes and the Brodmann areas. The captions indicate general functional specializations. The top view is a lateral view (from the side), whereas the lower one is a medial view (down the middle).

(Reproduced courtesy of www.appliedneuroscience.com .)

Remarkably, Brodmann ' s classification of cortical regions also serves as a basis for classifying cortical regions according to their function. Here are some examples.

- The primary visual cortex, also known as *area V1* , is the point of arrival for information from the retina. In anatomical terms, it is Brodmann area 17.
- Somatosensory information about the body gained through touch and body sense arrives in a region of the postcentral gyrus known as the *primary somatosensory cortex* . This is Brodmann area 3.
- We have already mentioned the primary motor cortex (the precentral gyrus). This is Brodmann area 4.

Even from an anatomical point of view, identifying segregated and distinct cortical regions can only be part of the story. We also need to know how the cortical regions are connected with each other. This would give an anatomical wiring diagram of the brain — or to use the standard terminology, a map of *anatomical connectivity* .

Exploring anatomical connectivity requires a whole new set of techniques. For example, *tract tracing* involves injecting a marker chemical into a particular brain region. Typical markers are radioactive amino acids or chemicals such as horseradish peroxidase (HRP). When the marker is injected near to the body of a nerve cell, it is absorbed by the cell body and then transported along the cell ' s axon. Looking to see where the marker ends up allows neuroanatomists to identify where the cell projects to — and doing this for enough cells allows them to work out the connections between different brain regions.

Tract tracing is an *invasive* technique. It is only possible to discover where HRP has been transported to by examining sections of the cortex through a microscope. This cannot be done on living creatures. And so neuroanatomists have primarily worked on the brains of nonhuman animals — primarily macaque monkeys, rats, and cats. Their results are often represented using *connectivity matrices* .

Figure 9.3 is an example from a very influential set of data on the visual system of the macaque monkey published in 1991 by Daniel J. Felleman and David Van Essen. The brain regions are abbreviated in a standard way. We can read off the matrix the regions to which any given region projects. Find the region of interest in the first column and then work your way across. If there is a " 1 " in the column corresponding to another brain region, then there is a connection going from the first to the second. If there is a " 0, " then no connection has been found. The gaps in the matrix indicate a lack of information.

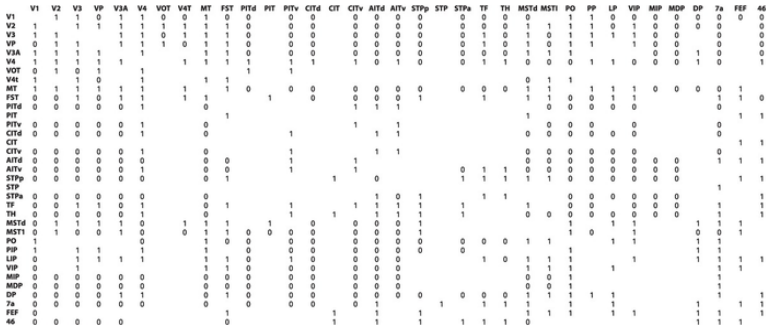


Figure 9.3 A connectivity matrix for the visual system of the macaque monkey.

(Adapted from Felleman and Van Essen 1991 .)

The same data can be presented in a form that makes it look much more like a wiring diagram. We see this in Figure 9.4 . The wiring-diagram format

makes it a little easier to visualize what is going on, but it doesn ′ t give quite as much information as the connectivity matrix.

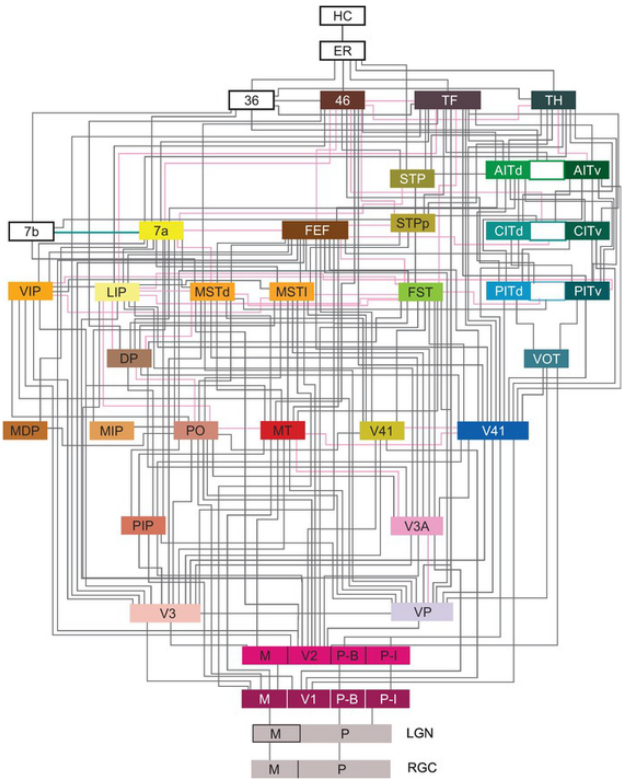


Figure 9.4 An anatomical wiring diagram of the visual system of the macaque monkey.

(Adapted from Felleman and Van Essen 1991 .)

Exercise 9.1 What type of information about anatomical connectivity do we get from a connectivity matrix but not from a wiring diagram?

Unfortunately, there are limits to what we can learn from anatomical connectivity.

- Anatomical connectivity data are largely derived from animal studies, whereas the brains that we are really interested in are our own. The information that we have about anatomical connectivity specifically in humans is largely derived from postmortem studies of human brains. Techniques for studying human anatomical connectivity in vivo, such as *diffusion tractography* , are being developed but are still in their infancy.
- Anatomical wiring diagrams do not carry any information about the direction of information flow between and across neural regions. There are typically at least as many *feedback* connections as *feedforward* connections.
- An anatomical wiring diagram tells us which brain regions are, in principle, able to " talk " directly to each other. But it does not tell us anything about how different brain regions might form circuits or networks to perform particular information-processing tasks.

9.2 Studying Cognitive Functioning: Techniques from Neuroscience

In addition to the principle of segregation, most neuroscientists also accept a *principle of integration* . This is the idea that cognitive functioning involves the coordinated activity of networks of different brain areas, with different types of tasks recruiting different networks of brain areas. So, we need to supplement information about anatomical connectivity with information about what actually goes on in the brain when it is performing specific

cognitive tasks. Neuroscientists have developed a number of techniques for doing this.

Mapping the Brain ′ s Electrical Activity: EEG and MEG

When neurons fire, they send electrical impulses down their axons. These *action potentials* are transmitted to the dendrites of other neurons at *synapses* . This electrical activity is a good index of activity in neurons. What neurons do is fire, and when they fire, they generate electricity. This activity can be measured in a number of different ways.

Neurophysiologists can record the discharge of action potentials in individual neurons by placing a microelectrode close to the cell being recorded. (For an illustration, see Figure 3.13 .) This technique has been used to identify neurons that are sensitive to particular stimuli. The discovery of what are known as *mirror neurons* is a very good example. A group of neuroscientists led by Giacomo Rizzolatti in Parma, Italy, identified neurons in monkeys that fire both when the monkey performs a specific action and when it observes that action being performed by an observer. This is illustrated in Figure 9.5 .

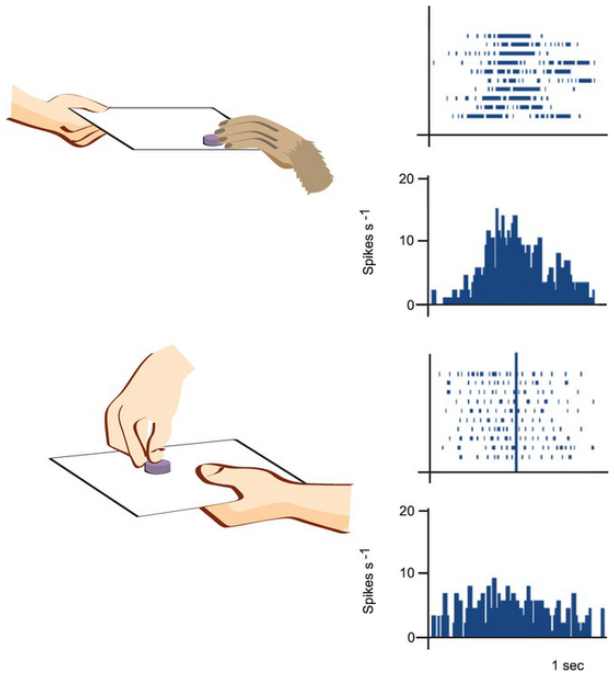


Figure 9.5 The results of single-neuron recordings of a mirror neuron in area F5 of the macaque inferior frontal cortex. The neuron fires both when the monkey grasps food (*top*) and when the monkey observes the experimenter grasping the food (*bottom*). Each horizontal line in the top diagram represents a single trial, and each nick represents the firing of a neuron. Neural activity is summed over trials in the two histograms.

(Adapted from Iacoboni and Dapretto 2006 .)

The activity of larger populations of neurons can be studied via human EEG, which requires little complicated machinery or disturbance to the subject. It uses electrodes attached to the skull and wired up to a computer. Each electrode is sensitive to the electrical activity of thousands of neurons, with the neurons nearest the electrode making the largest contribution to the output signal.

The coordinated activity of these neural populations can be seen in EEG as oscillatory waves at different frequencies. These frequencies are

typically labeled in terms of bands. The bands are named with letters from the Greek alphabet — from alpha through gamma. Confusingly, the alpha band is neither the lowest frequency nor the highest. The lowest-frequency activity takes place in the delta band. Delta-band activity is seen in very deep sleep (sometimes called *slow-wave sleep*).

In fact, different stages in the sleep cycle are associated with activity in different bands — and sleep specialists use EEG to identify and study sleep disorders. EEG can be used for other forms of medical diagnosis. So, for example, epilepsy is associated with a distinctive, " spiky " wave, as can be seen in Figure 9.6 .

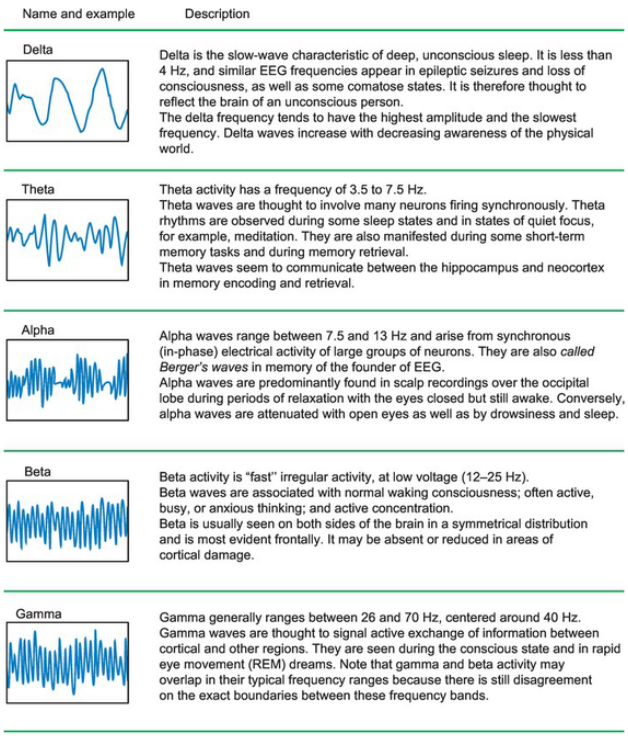


Figure 9.6 Typical patterns of EEG waves, together with where/when they are typically found.

(From Baars and Gage 2012 .)

EEG scans are particularly important because they give a reliable way of measuring what are known as *event-related potentials* (ERPs). An ERP is the electrical activity provoked by a specific stimulus. ERPs have a very fine temporal resolution. In other words, they are sensitive to very small differences in elapsed time.

The other principal technology for studying the electrical activity of large populations of neurons is *magnetoencephalography* (MEG). MEG measures the same electrical currents as electroencephalography, but it uses the magnetic fields that they produce. This allows a finer spatial resolution than EEG. It is also much less susceptible to distortion due to the skull than an EEG scan. But on the other hand, it brings with it all sorts of technical issues. For example, it can only be carried out in a room specially constructed to block all alien magnetic influences, including the earth ' s magnetic field. MEG is relatively little used in research neuroscience (as opposed to medical diagnosis) .

Mapping the Brain ' s Blood Flow and Blood Oxygen Levels: PET and fMRI

We turn now to PET and fMRI, which were introduced in Sections 3.4 and 3.5 , respectively. Instead of measuring electrical activity, these two techniques measure blood flow; more blood flows to a particular brain region when it is active.

PET measures blood flow directly by tracking the movement of radioactive water, whereas fMRI uses the indirect measure of blood oxygen levels in particular brain regions. This works because the increased neural activity in those regions does not consume all of the oxygen in the blood that

reaches them. So, the ratio of oxyhemoglobin to deoxyhemoglobin increases in areas that see increased blood flow. This gives rise to the BOLD signal.

PET and fMRI are both much more sensitive to spatial change and variation than to change and variation over time. In this respect, they are very different from EEG and MEG. Neuroimaging is a much better guide to how cognitive activity is distributed across the brain over a period of time than to the precise sequence of events as information is processed.

For this reason, functional neuroimaging is standardly used to identify networks of neural areas involved in carrying out cognitive tasks of a particular kind — those exploiting short-term memory, for example. This does not require a particularly fine temporal resolution. It simply requires being able to identify which neural regions are simultaneously active when the task is being performed. But the spatial resolution must be sufficiently fine-grained for the results to be interpretable in terms of standard anatomical maps of the brain — for example, to pinpoint activity in the premotor cortex (Brodmann area 6) or the orbitofrontal cortex (Brodmann area 11).

Table 9.1 summarizes some of the key features of these different techniques. As it shows, there is no technique for studying larger populations of neurons that has both a high temporal resolution and a high spatial resolution. Nonetheless, neuroscientists have combined techniques in order to gain a more comprehensive perspective . We will illustrate this through two case studies.

Table 9.1 Comparing techniques for studying connectivity in the brain

	DIRECTLY MEASURES	TEMPORAL RESOLUTION	SPATIAL RESOLUTION
Single-unit recording	Potentials in individual neurons and very small populations of neurons	High	High
EEG (electroencephalography)	Electrical activity of larger populations of neurons	High	Low
MEG (magnetoencephalography)	Magnetic fields produced by electrical activity of larger populations of neurons	High	Low
PET (positron emission tomography)	Cerebral blood flow in particular brain regions	Low	High
fMRI (functional magnetic resonance imaging)	Levels of blood oxygen in particular brain regions	Low	High

9.3 Combining Resources I: The Locus-of-Selection Problem

Perception is highly selective. At any given moment, we effectively ignore a huge amount of the information that our perceptual systems give us. We saw an example of this in Chapter 1 — the so-called *cocktail party phenomenon* . The same holds for vision. In principle, we can see things that are more or less level with our ears. Yet we are barely aware of much of our so-called *peripheral vision* . It is only when something in the periphery " catches our eye " that we realize quite how far our field of vision extends.

This selectivity is a very basic feature of perception. We only *focus on* or *attend to* a small proportion of what we actually see, hear, touch, and so on. Psychologists label the mechanism responsible for this very general phenomenon *attention* .

In Chapter 1 , we looked at Donald Broadbent ' s model of attention, in which attention functions as a selective filter. What the filter lets through depends on what the cognitive system as a whole is trying to achieve. In a cocktail party situation, for example, the filter might be tuned to the sound of a particular individual ' s voice.

In Broadbent ' s model, attention does its screening relatively early on in perceptual processing. For example, the selective filter screens out all the sounds that don ' t correspond to the voice of the person I am talking to long before my auditory systems get to work on parsing the sounds into words and then working out what is being said. His model is what is known as an *early-selection model* .

Other models claim that attention operates at a much later stage. These are *late-selection models* . In vision, for example, late-selection models hold that attention only comes into play once representations of sensory features (such as color, shape, and so on) have already been combined into representations of objects and those objects have been identified. The late-selection approach is taken, for example, in the object-based model of attention developed by cognitive psychologist John Duncan in the 1980s.

Exercise 9.2 Explain in your own words the difference between the late-selection and early-selection models of attention.

The *locus-of-selection problem* is the problem of determining whether attention is an early-selection phenomenon or a late-selection phenomenon. To solve it, we need a way of tracking perceptual information processing to identify when attention comes into play.

Combining ERPs and Single-Unit Recordings

This is, at bottom, a problem about the temporal organization of information processing: Does the processing associated with selective attention take place before or after the processing associated with object recognition? This

suggests using EEG to measure the ERPs evoked by visual information processing. EEG scans have a very high temporal resolution, sensitive at the level of milliseconds (thousandths of a second).

Remember that EEG (electroencephalography) is the general technique, whereas an ERP (the evoked reaction potential) is what the technique actually measures when it is *time-locked* with the onset of a particular stimulus. What we get from an ERP experiment is a wave that measures the electrical activity in the period of time immediately following the onset of the stimulus. The time is standardly measured in milliseconds, and the electrical activity is measured in microvolts (millionths of a volt).

We see a typical example in Figure 9.7 . The graph displays the ERP as a number of spikes and troughs. These are known as the *components* of the ERP and represent voltage deflections. The voltage deflections are calculated relative to a prestimulus baseline of electrical activity — which might, for example, be derived by measuring electrical activity at the tip of the nose.

WARNING: Please bear in mind a very confusing feature of ERP graphs. The *y* -axis represents negative activations above positive ones. This is very counterintuitive because it means that when the line goes up, the electrical activity is actually going down! And vice versa.

The time that elapses between stimulus onset and a particular component is known as that component ' s *latency* . The components of the ERP for vision have been well studied. The earliest is known as the *C1 component* . It is a negative component and appears at 50 – 90 ms after the appearance of the stimulus. There is a standard labeling system for subsequent components. These are labeled either *P* or *N* , depending on whether they are positive or negative. And they are given a number, which represents either their position in the ERP or their latency. The P1

component, for example, is the first positive component, whereas the P300 is a positive component that occurs 300 ms (i.e., 0.3 seconds) after the stimulus is detected.

The P300 typically occurs in response to unexpected or novel stimuli. It is often taken as a sign that higher cognitive processes, such as attention, are involved in processing the stimulus. The graph in Figure 9.7b has the C1, N1, and P1 components marked. It is also possible to see the P200 component and a (slightly delayed) P300.

The ERP wave displays an *attention effect* . Certain components of the wave change if the subject is attending to the stimulus. Figure 9.7a illustrates the type of experiment used to elicit the attention effect. The subject is asked to attend to one of two boxes on a screen. Stimuli are presented at various places on the screen, and the ERPs are measured both for the case where the stimulus is in the box being attended to and the case where it is elsewhere.

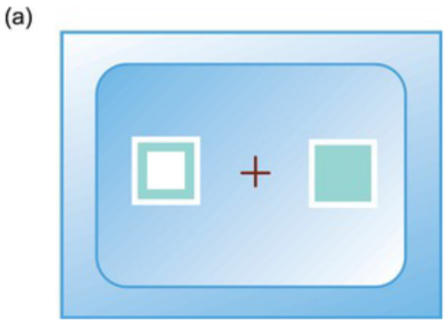


Figure 9.7a Common experimental design for neurophysiological studies of attention. The outline squares are continuously present and mark the two locations at which the solid square can be flashed.

(Courtesy Stephen J. Luck and Michelle A. Ford.)

Figure 9.7b illustrates the results of a typical experiment. The solid line shows the ERP when subjects are attending and the dotted line when they are

not attending. There are important differences in two of the components — together with an important nondifference in one component. The nondifference first: there is no significant difference in the C1 component between the attended and the unattended cases. But there are significant differences in the P1 and N1 components. The P1 component is the first significant positive component, and the N1 is the first significant negative component. Both are larger when the subject is attending to the box in which the stimulus appears.

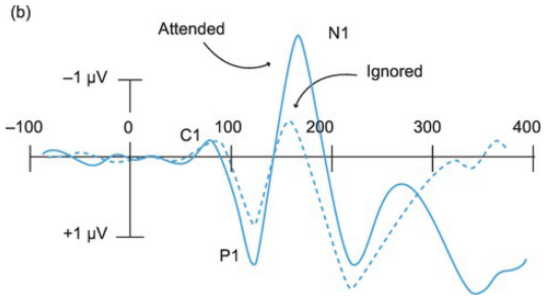


Figure 9.7b Example of the occipital ERPs recorded in a paradigm of this nature. Note that the C1 wave (generated in area V1) shows no attention effect, whereas the P1 and the N1 waves (generated in the extrastriate cortex) are larger for the attended stimuli.

(Courtesy Stephen J. Luck and Michelle A. Ford.)

This suggests that there are two additional bursts of information processing taking place roughly 100 and 200 ms after stimulus onset when attention is exercised. But how does this help us to decide whether attention is an early-selection phenomenon or a late-selection phenomenon?

The ERP data on their own cannot settle this question. To make progress, we need to triangulate ERP data with what we know about what goes on in different parts of the brain.

As observed in Section 3.2 , neurophysiologists generally accept that (at least in the visual system of the macaque monkey) object identification exploits the so-called *ventral pathway* that begins in V1 (the *striate cortex*) and then progresses through areas V2 and V4 en route to the inferotemporal cortex. Of these areas, V1 is responsible for processing basic shape. Visual areas V2 and V4 (which are *extrastriate* areas) are thought to process more advanced information about shape, together with information about color, texture, and so on.

The different areas in the object-identification pathway process different types of information separately but in parallel. Moreover, the information processing in V1, V2, and V4 takes place *upstream* of any integration of these different types of information. In other words, everything in early visual areas such as V1, V2, and V4 takes place *before* the visual system is working with representations of objects.

This gives a clear criterion for thinking about the locus-of-selection problem. Because attention is a late-selection phenomenon only if it comes into play when the visual system has generated (and perhaps identified) representations of objects, any evidence that the exercise of attention affects processing in the early visual areas will be evidence that attention is an early-selection phenomenon.

There is a range of evidence connecting different components of the ERP wave to processing in different visual areas. The C1 component, for example, is thought to reflect processing in the striate cortex (V1). Because the C1 component is constant across both the attended and the unattended conditions, we can conclude that processing in V1 is not modulated by attention. On the other hand, there is evidence connecting the P1 and N1 components with processing in the extrastriate cortex (i.e., in areas such as

V2 and V4). So, the ERP data do seem to show that attention affects early visual processing, which supports the early-selection view rather than the late-selection view.

This is reinforced by single-unit recordings. Figures 9.7c and 9.7d show recordings in V1 and V4 while monkeys are performing a task similar to that depicted in Figure 9.7a . As the graphs show, there is no difference between levels of activity in V1 across the attended and unattended conditions. But there are significant differences in V4. This is certainly consistent with the hypothesis that attention is an early-selection phenomenon.

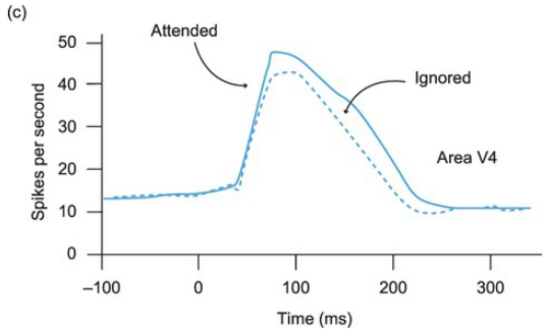


Figure 9.7c Single-unit responses from area V4 in a similar paradigm. Note that the response is larger for attended compared with ignored stimuli.

(Courtesy Stephen J. Luck and Michelle A. Ford.)

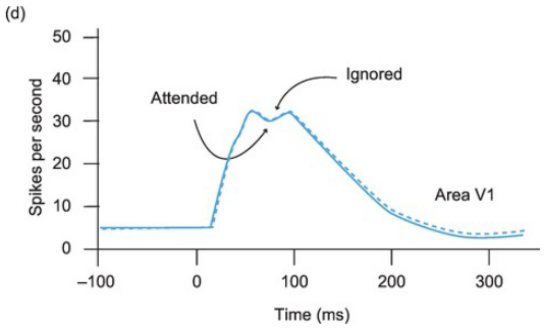


Figure 9.7d Single-unit responses from area V1 showing no effect of attention.

(Courtesy Stephen J. Luck and Michelle A. Ford.)

There is a clear " take-home message " here. Although there are no techniques or technologies for studying cognitive activity directly, and although each of the techniques has significant limitations, we can overcome many of the limitations by combining and triangulating the different techniques. The high temporal resolution of EEG complements the high spatial resolution of imaging technologies such as PET. And predictions from studies of humans using these techniques can be calibrated with electrophysiological studies on monkeys .

9.4 Combining Resources II: Networks for Attention

Attention raises other important questions. For example:

- Which brain areas are involved in attention?
- How is attention related to other cognitive processes, such as memory and action planning?

- How does the brain direct attention to particular objects and particular places?

Exploring these questions in this section will illustrate the power of functional neuroimaging — and also, to continue one of the themes of this chapter, how neuroimaging data can be calibrated with electrophysiological experiments.

There are different varieties of attention. We can attend to one object among others — to the unfamiliar bird in the flock of sparrows, for example. Or we can attend to one part of an object rather than another — to the bird ′ s head or beak rather than its wings. Alternatively, we can attend to places — to the place where we expect the bird to fly to next.

The experiments discussed in the previous section focused on the last of these types of visual attention. Subjects were asked to focus on a particular location on the screen (marked by a box) — a location at which a stimulus might or might not appear. And it is what we will be focusing on in this section also. Neuroscientists and psychologists term it *spatially selective attention* (or *visuospatial attention*).

So, which brain areas are involved in spatially selective attention? Studies in the 1990s identified a network of cortical areas implicated in visuospatial attention. The specific tasks varied, but they all involved subjects directing attention to stimuli in the periphery of their visual field without moving their eyes. This is very important. Typically, we attend to different objects in the visual field by making very quick (and unconscious) eye movements known as *saccadic eye movements* . Experimenters studying visuospatial attention, however, are interested in attention as a mechanism that operates independently of eye movements — a mechanism that can be

directed at different peripheral areas while gaze is fixated on a central point. Researchers call this *covert attention* .

Figure 9.8 summarizes a number of these studies. It identifies a network of frontal and parietal areas active when subjects direct covert attention to peripheral areas in the visual field. The existence of this frontoparietal cortical network is widely accepted among researchers of attention and has been confirmed by retrospective analyses of PET and fMRI data.

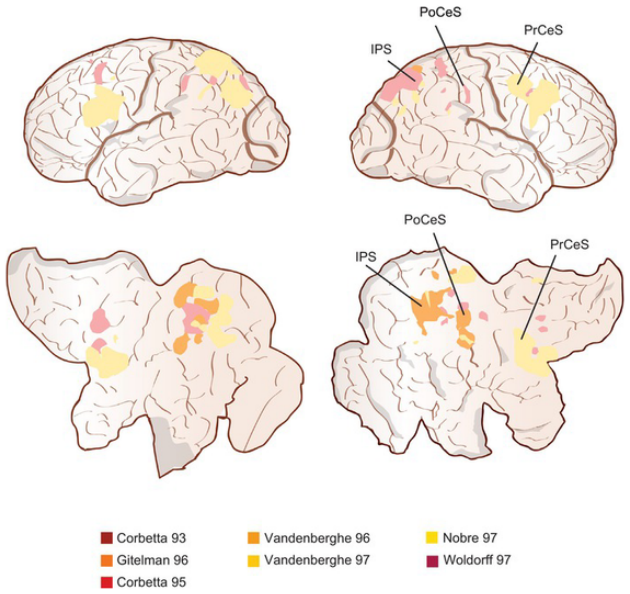


Figure 9.8 Frontoparietal cortical network during peripheral visual attention. Common regions of activation across studies include the intraparietal (IPS), postcentral (PoCeS), and precentral sulcus (PrCeS).

(Adapted from Gazzaniga 2000 .)

But simply identifying a network of brain areas involved in visuospatial attention does not in itself tell us much about how attention works. It does not tell us about how attention is related to other cognitive processes, such as memory or action planning. And it does not tell us anything about how

exactly the brain directs attention to particular locations in space. We turn to those questions now.

Two Hypotheses about Visuospatial Attention

There are two dominant hypotheses about how visuospatial attention works.

The first hypothesis is that visuospatial attention exploits memory mechanisms — to attend to a specific location, we need to remember that location actively. If this is right, then we would expect brain networks associated with spatial working memory to be active during tasks that involve attention.

An alternative hypothesis is that directing attention to a particular location is closely connected to preparing to move to that location — even in the case of covert attention, where the intention to move is the intention to move the eyes. This predicts that brain areas associated with motor planning will be active in tasks that exploit visuospatial attention.

The two hypotheses are not necessarily exclusive (fortunately, because there is considerable experimental support for both).

Some of the evidence comes from single-neuron studies on monkeys. Carol Colby and her collaborators made recordings from an area in the parietal cortex known as the *lateral intraparietal area* (LIP) while monkeys were carrying out a *delayed-saccade task* . The LIP (which we looked at in a different context in Section 7.3) is thought to play an important role in short-term memory of spatial locations.

In an ordinary saccade task, monkeys are trained to make a saccade (i.e., quickly move both eyes) from a central fixation point to a stimulus as soon as the stimulus appears. In a delayed- saccade task, the monkeys are

trained not to make the saccade until the fixation point disappears — by which time the stimulus has disappeared (see Figure 9.9). When the fixation point disappears, they then have to make a saccade to the location where the stimulus originally appeared.

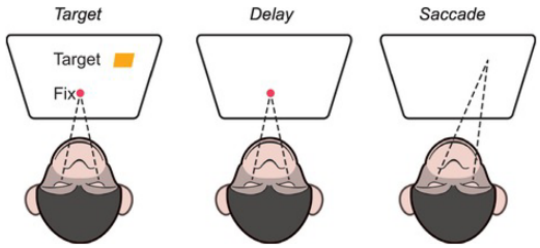


Figure 9.9 An illustration of a typical delayed-saccade task. The monkeys are trained to withhold their saccade to the visual target until the fixation point disappears. Note that the head does not move during the task.

(From Gazzangia 2000.)

Success on the delayed-saccade task requires the monkeys to remember where the stimulus appeared if they are to make a successful saccade. This short-term memory of spatial location is typically called *spatial working memory* .

It turns out that the firing rates of neurons in the LIP go up both when monkeys are performing delayed-saccade tasks (and so exercising spatial working memory) and when they are carrying out peripheral attention tasks, such as those discussed in the previous section.

This electrophysiological evidence is backed up by neuroimaging studies on humans. Both PET and fMRI studies have shown a significant overlap between the brain areas activated in visuospatial attention tasks and those active during tasks that require subjects to use working memory to

store and manipulate information about spatial locations — as depicted in the two diagrams on the left-hand side in Figure 9.10 .

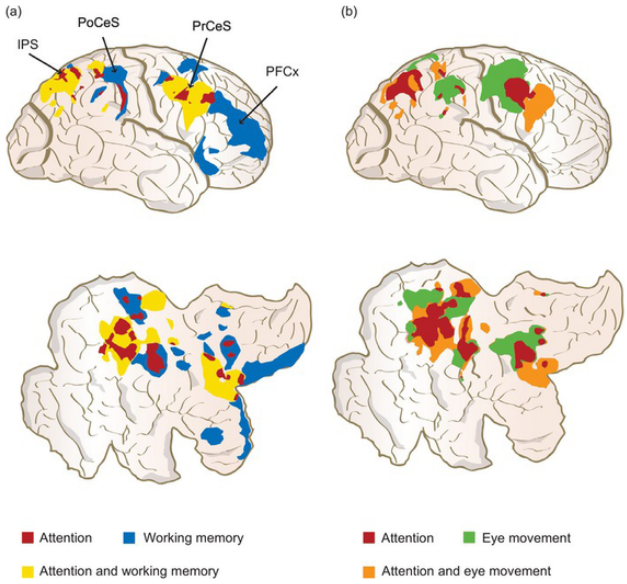


Figure 9.10 (a) Peripheral attention versus spatial working memory versus saccadic eye movement across studies. Left: Regions active for peripheral attention (*red*), regions active for spatial working memory (*blue*), and regions of overlap (*yellow*). Note the striking overlap in the parietal cortex, the partial overlap in the precentral region, and the exclusive activation of the prefrontal cortex (PFCx) for spatial working memory. (b) Comparison between peripheral attention (*red*) and saccadic eye movements (*green*). Note the strong overlap (*orange*) in both the parietal and precentral regions. There is no activation in the prefrontal cortex.

(Adapted from Gazzaniga 2000 .)

These diagrams show that although there seem to be separate cortical networks for visuospatial attention and spatial working memory, these networks overlap very significantly in the parietal cortex. This is highly consistent with the electrophysiological results.

Turning now to the relation between visuospatial attention and preparatory motor responses, the two diagrams on the right-hand side of Figure 9.10 analyze experiments on the relation between covert attention and saccadic eye movements. For example, Maurizio Corbetta of Washington University in St. Louis scanned subjects both during conditions that required them to shift attention while maintaining their gaze fixed on a fixation point and during conditions in which gaze and attention shifted simultaneously.

As the diagrams show, there is significant overlap across the covert attention and the saccadic eye movement tasks, both in the parietal and in the precentral region (where the overlap is much stronger than in the working-memory experiments).

This discussion of visuospatial attention illustrates an important methodological lesson. Progress in this area depends on combining and calibrating what is learned from each of these techniques. We do not have any *direct* measures of cognitive activities such as visuospatial attention. But we do have the next best thing, which is a wide range of *indirect* measures. Single-unit recordings, PET, fMRI, and EEG all give us very different perspectives on visuospatial attention. We can use some techniques to compensate for the weaknesses of others. And we have powerful tools for cross-checking and integrating information from different sources. We have seen how this works in the case of visuospatial attention. This is an excellent case study in how neuroscientists are moving toward the goal of providing a cognitive wiring diagram of the brain.

9.5 From Data to Maps: Problems and

Pitfalls

Neuroimaging yields unparalleled insight into the structure and organization of the mind — perhaps more so than anything else in the neuroscientist ' s toolkit. But it needs to be used with caution. It is easy to be seduced by the brightly colored images that emerge from software packages for interpreting neuroimaging data. These images look very much like maps of the brain in the process of thinking. And so it is easy to think that neuroimaging gives us a " window on the mind. " In this section, we will see why we need to be much more cautious.

From Blood Flow to Cognition?

Neuroimaging only measures cognitive activity indirectly. fMRI measures the BOLD signal, whereas PET measures cerebral blood flow. But as we saw in Section 3.6 , there is a lively debate about the neural correlates of the BOLD signal. Researchers are calibrating fMRI data with electrophysiological techniques in order to try to work out whether the BOLD signal is correlated with the firing rates of populations of neurons or whether it is correlated with the local field potentials (which are thought to reflect the inputs to neurons, rather than their outputs). We looked at some experimental evidence (from Logothetis and his collaborators) that seems to point to the second possibility.

But even if we had a conclusive answer to this question, we would still be a long way from a clear picture of the relation between variation in the BOLD signal and information processing in the brain. This is because we do not have any generally accepted models of how populations of neurons process information in particular brain areas. One illustration is that the

BOLD signal gives us no indication of whether the activity it measures is excitatory or inhibitory.

Noise in the System?

Neuroimaging ′ s great strength is its spatial resolution, but this comes at a cost.

The basic spatial unit in fMRI is the *voxel* . This is a three-dimensional version of a pixel (the name is a combination of the words *volume* and *pixel*). The basic unit of data obtained from fMRI is the BOLD signal in each voxel. The spatial resolution is directly correlated with the size of the voxels — the smaller the voxel, the higher the spatial resolution. The problem, though, is that the strength of the signal is directly correlated with the size of the voxel — the smaller the voxel, the lower the signal strength.

For some brain areas, particularly those involving basic perceptual processing or simple motor behaviors (such as finger tapping), experimenters can design tasks that elicit strong signals even when the voxel size is small. Things are not so straightforward, however, for more complex types of processing — particularly those performed by distributed networks of neural areas. In this case, it is often necessary to increase the voxel size in order to capture smaller fluctuations in the BOLD signal. Unsurprisingly, this decreases the spatial resolution. But it also has a less expected consequence.

Increasing the voxel size increases the range of different types of brain tissue occurring in each voxel. When the voxel includes extraneous material, such as white matter or cerebrospinal fluid, this can distort the signal, giving rise to what is known as *partial-volume effects* . It can also happen that a

single voxel contains more than one cell type, whereas neuroimaging data are standardly interpreted on the tacit assumption that voxels are homogeneous.

There are other ways in which noise can get into the system. Because everybody ′ s brain is subtly different, meaningful comparisons across different subjects depend on *normalizing* the data — that is, by reinterpreting the data from each subject with the use of a brain atlas with a common coordinate system, or what is known as a *stereotactic map* . This requires very complicated statistical techniques, which may themselves introduce distortion in the data.

And there are many different brain atlases, such as the Talairach – Tournoux atlas, the Montreal Institute of Neurology (MNI) atlas, and the Population-Average, Landmark- and Surface-Based (PALS) atlas developed by David Van Essen at Washington University in St. Louis. Because different research groups often use a different atlas, this can make the business of comparing and contrasting different studies a tricky undertaking.

Functional Connectivity versus Effective Connectivity

Neuroimaging helps us to understand the *connectivity* of the brain. But the wiring diagram that we get from fMRI and PET does not tell us anything *directly* about how information flows through the brain.

A single experiment can tell us which brain areas are simultaneously active while subjects are performing a particular task, but it does not reveal, for example, the order in which the areas are active. The diagrams that present the results of neuroimaging experiments only show which areas " light up together. " They identify a network of areas that are simultaneously

active when certain tasks are performed. But they really only identify correlations between the activity levels of different brain areas.

For this reason, neuroimagers distinguish *functional connectivity* from *effective connectivity*. Functional connectivity is a statistical notion. It is standardly defined in terms of statistical correlations between levels of activity in physically separate parts of the brain.

But to understand how the brain processes information, we really need what neuroscientists call *effective connectivity*. This is a measure of how neural systems actually interact. Effective connectivity captures the idea that information processing is a causal process. Information flows through different brain areas in a particular order. What happens to the information at earlier stages affects how it is processed at later stages.

Exercise 9.3 Explain in your own words the distinctions between anatomical, functional, and effective connectivity.

Neuroimaging is much better at telling us about functional connectivity than about effective connectivity. PET and fMRI are tools specialized for studying correlation, not causation.

Admittedly, there are ways of deriving conclusions about effective connectivity from neuroimaging data. For example, one can design a series of experiments in a way that yields information about the flow of information. We looked at a very nice example of this back in Section 3.4. Steve Petersen and his collaborators drew significant conclusions about the stages of lexical processing from a series of PET experiments using the

paired-subtraction paradigm. Their model is a model of effective connectivity in this domain.

> **Exercise 9.4** Look back at the lexical processing experiments described in Section 3.4 and explain in your own words how their experimental design overcomes some of the problems raised by the distinction between functional and effective connectivity.

And as I have been stressing throughout this chapter, neuroimaging results can always be calibrated and triangulated with other tools and techniques, such as EEG and electrophysiology. The locus-of-selection problem shows how data from neuroimaging, EEG, and electrophysiology can be combined to develop a model of the effective connectivity of covert attention.

Nonetheless, we do have to be careful in how we interpret the results of neuroimaging experiments. We must be very careful not to draw conclusions about effective connectivity (causal relations between brain areas and how information flows between them) from data that only tell us about functional connectivity (i.e., correlations between BOLD signal levels in those areas).

Summary

This chapter has looked at how cognitive neuroscience can help us to construct a wiring diagram for the mind. We began by highlighting the complex relations between functional structure and anatomical structure in the brain and then looked at some of the techniques for tracing anatomical

connections between different brain areas. Completely different tools are required to move from anatomical connectivity to functional connectivity. We looked at various techniques for mapping the brain through the measurement of electrical activity and blood flow and blood oxygen levels. These techniques all operate at different degrees of temporal and spatial resolution. As we saw in two case studies, each having to do with a different aspect of the complex phenomenon of attention, mapping the functional structure of the brain requires combining and calibrating different techniques. At the end of the chapter, we reviewed some of the pitfalls in interpreting neuroimaging data.

Checklists

It is a basic principle of neuroscience that the cerebral cortex is divided into segregated areas with distinct neuronal populations (the *principle of segregation*).

1. These different regions are distinguished in terms of the types of cells they contain and the density of those cells. This can be studied using staining techniques.

2. Neuroscientists can study *anatomical connectivity* (i.e., develop an anatomical wiring diagram of the brain) by using techniques such as tract tracing or diffusion tractography.

3. Most of the evidence comes from animal studies. Neuroscientists have developed well-worked-out models of anatomical connectivity in macaque monkeys, rats, and cats.

4. Anatomical connectivity is not necessarily a guide to functional connectivity.

Neuroscientists also adopt the *principle of integration* — that cognitive functioning involves the coordinated activity of networks of different brain areas.

1. Identifying these networks requires going beyond anatomical activity by studying what goes on in the brain when it is performing particular tasks.

2. Some of the techniques for studying the organization of the mind focus on the brain ′s electrical activity. These include electrophysiology, electroencephalography (EEG), and magnetoencephalography (MEG).

3. These techniques all have high temporal resolution — particularly EEG when it is used to measure event-related potentials (ERPs). But the spatial resolution is lower (except for electrophysiology using microelectrodes).

4. Other techniques measure blood flow (positron emission tomography [PET]) and levels of blood oxygen (functional magnetic resonance imaging [fMRI]). These techniques have a high spatial resolution but lower temporal resolution.

The *locus-of-selection* problem is the problem of determining whether attention operates early in perceptual processing or upon

representations of objects. It provides a good illustration of how neuroscientists can combine different techniques.

1. The problem has been studied using EEG to measure ERPs. Attentional effects appear relatively early in the ERP wave following the presentation of a visual stimulus.

2. These results can be calibrated with PET studies mapping stages in the ERP wave onto processing in particular brain areas. This calibration reveals attentional effects in areas such as V2 and V4, which carry out very basic processing of perceptual features.

3. This resolution of the locus-of-selection problem seems to be confirmed by single-unit recordings in monkeys.

Neuroimaging techniques can help identify the neural circuits responsible for attention.

1. Preliminary evidence from patients with brain damage (e.g., with hemispatial neglect) points to the involvement of frontal and parietal areas in visuospatial attention.

2. This has been confirmed by many experiments on covert attention using PET and fMRI.

3. PET and fMRI experiments on humans, together with single-neuron experiments on monkeys, have shown that tasks involving visuospatial attention also generate activation in brain networks responsible for planning motor behavior and for spatial working memory.

Although neuroimaging is a powerful tool, it is not a " window on the mind, " and neuroimaging data should be interpreted with

caution.

1. Neuroimaging techniques can only measure cognitive activity indirectly. PET measures blood flow, and fMRI measures the blood-oxygen-level – dependent (BOLD) signal. There is a controversy in neuroscience about what type of neural activity is correlated with the BOLD signal (see Section 3.6) — and there is no worked-out theory about how that neural activity functions to process information.

2. There are many opportunities for noise to get into the system in neuroimaging experiments. Partial-volume effects can occur when the voxel size is large, and distortions can occur when data are being normalized to allow comparison across subjects.

3. Neuroimaging techniques are much better at telling us about *functional connectivity* (correlations between activation levels in different brain areas as a task is performed) than about *effective connectivity* (how information flows between different brain areas and how they influence each other).

Further Reading

The explosion of interest in cognitive neuroscience in the last couple of decades has generated a huge literature. For keeping up to date with contemporary research, the journal *Trends in Cognitive Sciences* regularly contains accessible survey articles. Authoritative review articles on most of the key topics studied by cognitive neuroscientists can be found in the fifth edition of *The Cognitive Neurosciences* , edited by Michael Gazzaniga and George Mangun (2014). The four earlier editions (Gazzaniga 1995 , 2000 ,

2004 , 2009) also contain much useful material. Gazzaniga is one of the authors of an influential textbook on cognitive neuroscience (Gazzaniga et al. 2013). Chapter 3 is a useful introduction to the methods of cognitive neuroscience. Also see Baars and Gage 2010 .

Zeki 1978 was one of the first papers to identify functional specialization in the primate visual system. David Van Essen ' s work is accessibly presented in Van Essen and Gallant 1994 . The much-cited paper discussed in the text is Felleman and Van Essen 1991 . Reviews of other classic work can be found in Colby and Goldberg 1999 and Melcher and Colby 2008 . Orban et al. 2004 is an interesting discussion of the challenges in comparing the neurobiology of cognitive function across humans and macaque monkeys. Also see Passingham 2009 . Ji et al. 2019 presents a large-scale functional map of the human brain, with an accompanying network atlas.

An interesting trend in recent discussions of anatomical connectivity has been the use of mathematical tools from graph theory — in particular the idea of small-world networks. There is a very useful introduction in Bassett and Bullmore 2006 . See Minati et al. 2013 for a more up-to-date review. Jirsa and McIntosh 2007 is a collection of surveys of different aspects of neural connectivity. For article-length surveys, see Ramnani et al. 2004 , Bullmore and Sporns 2009 , and Friston 2011 . Bressler et al. 2008 uses Granger causality to explore effective connectivity in the neural basis of visual – spatial attention. For more on Granger causality, see Deshpande and Hu 2012 and Friston et al. 2013 .

For a spirited methodological defense of MEG, see Baillet 2017 , and Schomer and Lopes da Silva 2020 for the latest edition of the leading textbook on EEG. Abreu et al. 2018 reviews techniques for integrating EEG

and fMRI. There has been much discussion of the pitfalls and advantages of using neuroimaging techniques to study cognitive function in the human mind. In addition to research on the neural basis of the BOLD signal discussed in Chapter 3 (see the references there), researchers have focused on the methodology of inferring cognitive function from selective patterns of activation. See, for example, Henson 2006 and Poldrack 2006 . For a review of the state of fMRI at the time from a leading researcher, see Logothetis 2008 . Also see Ashby 2011 , Charpac and Stefanovic 2012 , Machery 2012 , and Poldrack et al. 2011 . Op de Beeck and Nakatani 2019 is a recent textbook on human neuroimaging, and Poldrack 2018 is a book-length survey of its power and pitfalls. Poldrack et al. 2017 offers a cautious roadmap for the future of functional neuroimaging.

For surveys of research into selective attention, see Hopfinger et al. 2004 and Thigpen and Keil 2017 . The experimental work reported in Section 9.3 is described more fully in Luck and Ford 1998 . Stephen Luck is the author of an important textbook on ERP techniques (Luck 2014). The introductory chapter from the first edition can be downloaded from the online resources. See also his coedited volume, Luck and Kappenman 2011 . Awh et al. 2012 tries to underline the early selection – late selection dichotomy.

Humphreys et al. 1999 contains many useful papers on the psychology and neuroscience of attention, as does Posner 2004 . For more details of the findings discussed in Section 9.4 , see Chelazzi and Corbetta 2000 . Other good reviews on a wide variety of attention phenomena can be found in Chapters 8 and 10 of Baars and Gage 2010 , as well as in Carrasco 2011 , Chun et al. 2011 , Posner 2017 , and Carrasco 2018 . Howard et al. 2020 is a

recent study of the relation between spatial attention and spatial working memory .

Part III



Applications

Chapter 10

Models of Language Learning



Overview

10.1 Language and Rules

Understanding a Language and Learning a Language

10.2 Language Learning and the Language of Thought: Fodor's Argument

10.3 Language Learning in Neural Networks

The Challenge of Tense Learning

Connectionist Models of Tense Learning

10.4 Bayesian Language Learning

Probabilities in Word and Phrase Segmentation

Understanding Pronouns

Learning Linguistic Categories

Overview

Language is a highly sophisticated cognitive achievement. Without it, our cognitive, emotional, and social lives would be immeasurably impoverished. And it is a truly remarkable fact that almost all human children manage to

arrive at more or less the same level of linguistic comprehension and language use. Cognitive scientists have devoted an enormous amount of research to trying to understand how languages are learned. This chapter looks at language learning from three of the theoretical perspectives discussed in earlier chapters:

- The language-of-thought hypothesis (a version of the physical symbol systems hypothesis)
- Connectionist neural networks
- Probabilistic Bayesian models

Section 10.1 introduces some basic theoretical challenges in explaining how we understand and learn languages. Because language is a paradigmatically rule-governed activity, it can seem very plausible to conceptualize linguistic understanding as a matter of deploying linguistic rules. This raises the question of where knowledge of the rules comes from and how they are learned.

Section 10.2 looks at one answer to that question. According to Jerry Fodor, young children learn linguistic rules by a process of hypothesis formation and testing. This process is itself a linguistic activity, but Fodor argues that it cannot be carried out in a natural language. He thinks that it takes place in what he calls the *language of thought* , which he thinks is *innate* .

Section 10.3 explores the very different connectionist approach, which uses neural networks in which rules are not explicitly represented. We look at how neural networks can be trained to learn the past tense of English verbs, both regular and irregular. As we ′ ll see, the learning trajectory of these networks bears striking resemblances to the learning trajectory of human infants coming to terms with the complexities of verbs in English.

Like the connectionist approach, the Bayesian models considered in Section 10.4 explicitly set themselves against *nativist* or *innatist* accounts such as Fodor ' s. Bayesians think that arguments for innatism about language have significantly underestimated how much can be learned from statistical regularities and through applying Bayesian principles like those discussed in Chapter 7 .

10.1 Language and Rules

Every language is governed by grammatical rules. These rules, painfully familiar to anyone who has tried to learn a second language, govern how words can be put together to form meaningful sentences. But grammatical rules are only the tip of the iceberg. Linguists devote much of their time to trying to make explicit the much more fundamental rules that are supposed to apply to all languages, irrespective of their particular grammars.

Back in Section 1.3 , we looked briefly at the version of transformational grammar proposed by Noam Chomsky in the 1950s. Chomsky was proposing transformational rules by which a sentence with one type of grammatical structure could be transformed into a sentence with a different grammatical structure but a similar meaning — for example, " John has hit the ball " into " The ball has been hit by John. " Chomsky ' s view on what these rules actually are has changed many times over the years, but he has never abandoned the basic idea that the deep structure of language is governed by a body of basic rules.

The rule-governed nature of language makes thinking about language a very interesting test case for comparing and contrasting the different models of information processing that we have looked at, which all think about rules very

differently. As we saw in Chapter 4 , the physical symbol system hypothesis takes information processing to be manipulating physical symbol structures according to rules explicitly represented within the system. In contrast, in Chapter 5 , we learned that it is not really possible to distinguish rules and representations in artificial neural networks (apart from the algorithm that governs how the network updates its activation levels). Nor, as we saw in Chapter 7 , do Bayesian models explicitly encode rules (besides those required to manipulate probabilities and utilities).

We turn now to some of the theoretical reasons that have been given for thinking that the physical symbol system hypothesis (particularly in its language-of-thought incarnation) is the only way of making sense of the complex phenomenon of linguistic comprehension and language learning. And then, in the rest of the chapter, we will test the power of those arguments by looking at neural networks and Bayesian models of specific aspects of language learning .

Understanding a Language and Learning a Language

Understanding a language is partly a matter of understanding what individual words mean and partly a matter of understanding the rules governing how those words can be combined into sentences. What does this understanding consist of? The default hypothesis is that understanding a language is fundamentally a matter of mastering the relevant rules. This applies to the vocabulary of a language no less than to its grammar and deep structure. We can think of understanding the meaning of a word in terms of mastery of the rule that governs its application — the rule, for example, that the word *dog* refers to four-legged animals of the canine family and the rule that the word

square applies to four-sided shapes with sides of equal size and each corner at an angle of ninety degrees.

But everything depends on how we think about mastering a rule. At one extreme is the view that there is no more to mastering a linguistic rule than being able to use words in accordance with the rule. Competent language users need not *represent* the rule in any way. They just need to be able to distinguish applications of the word that fit the rule from applications that do not. This minimalist conception of linguistic understanding is associated with the Austrian philosopher Ludwig Wittgenstein. It makes linguistic understanding much more of a practical ability than a theoretical achievement.

For many cognitive scientists, though, this conception of rules is far too weak. After all, the rock that falls in accordance with Newton ′ s law of gravity cannot in any sense be said to have mastered that law. Mastering linguistic rules certainly requires using words in accordance with the rule, but it is not just a practical ability. That is why many theorists think that language users can only use words in accordance with linguistic rules if they represent those rules. Language users somehow manage to compare possible sentences with their internalized representations of the rules. This is the other extreme. It makes linguistic understanding, at bottom, a theoretical achievement.

Moreover, how one thinks about linguistic understanding has direct implications for how one thinks about language learning. In the next few sections, we will be looking at issues of *innatism* or *nativism* (terms that I ′ ll use interchangeably). Some accounts of linguistic understanding have the consequence that aspects of language must be innate because they seem, in principle, to be incapable of being learned by young children, given the limited resources and evidence that children have. The next section will look at one example of this type of theory, Fodor ′ s argument that language learning requires an innate language of thought. And then, in Sections 10.3 and 10.4 ,

we will look at two prominent anti-innatist approaches — connectionist models of language in Section 10.3 and then Bayesian approaches in Section 10.4 .

10.2 Language Learning and the Language of Thought: Fodor ' s Argument

This section examines an influential version of the physical symbol system hypothesis. This is the language-of-thought hypothesis, proposed by the philosopher Jerry Fodor and very influential in cognitive science.

Fodor holds a strong version of the rule-based conception of language learning. He thinks of the process of acquiring a language as a lengthy process of mastering the appropriate rules, starting with the simplest rules governing the meaning of everyday words, moving on to the simpler syntactic rules governing the formation of sentences, and then finally arriving at complex rules such as those allowing sentences to be embedded within further sentences and the complex transformational rules discussed by Chomsky and other theoretical linguists.

His argument from the rule-based conception of language learning to the language of thought is in his book *The Language of Thought* . It starts off from a particular way of thinking about the rules governing what words mean. According to Fodor, these are *truth rules* , so called because they spell out how words contribute to determining what it is for sentences in which they feature to be true. Mastering truth rules may not be all that there is to understanding a language. But Fodor is emphatic that we will not be able to understand a language without mastering truth rules. They may not be sufficient but are certainly necessary, he claims.

Consider the sentence " Felicia is tall. " This is what logicians call an *atomic sentence* . It is made up simply of a proper name (" Felicia ") and a predicate (" ___ is tall, " where the gap indicates that it needs to be " completed " by a name of some sort). Proper names are names of individuals, and predicates are names of properties. And so, this gives us a very straightforward way of thinking about what makes an atomic sentence such as " Felicia is tall " true. The sentence is true only if the individual named by the proper name (i.e., Felicia) does indeed have the property named by the predicate (i.e., the property of being tall). So, the atomic sentence " Felicia is tall " is true only if Felicia is tall. This is the sentence ' s *truth condition* (TC):

TC: " Felicia is tall " is true only if Felicia is tall.

According to Fodor, though, someone can only understand TC if they already understand the sentence " Felicia is tall " (because this very sentence features in the truth condition, both inside and outside quotation marks). But then the truth condition can only be intelligible to someone who already understands the expressions " Felicia " and " ___ is tall. " It cannot help us to make sense of how someone can *learn* to use those expressions.

So, Fodor thinks, we need something more than truth conditions, such as TC in the previous example, in order to make sense of linguistic comprehension and language learning. We need rules that will tell us which individual the name " Felicia " refers to and which property is named by the predicate " ___ is tall. " If these rules are to be learnable, then they must be stated in terms of expressions that the language user is already familiar with. In fact, we really need something like the following rule:

TC*: " Felicia is tall " is true only if X is G.

Here, *X* stands for another name for Felicia — one that the language user already understands (perhaps *X* might be "George's sister"). Likewise, *G* stands for another way of naming the property of being tall (perhaps *G* might be "greater than average in height"). This is what Fodor calls a *truth rule*.

Exercise 10.1 Explain in your own words the difference between the truth condition, TC, and the truth rule, TC*.

So, putting all this together, Fodor argues that learning a language has to involve learning truth rules. He thinks that this places some very fundamental constraints on any information-processing account of language learning. Learning a truth rule such as TC* is, he thinks, a matter of forming hypotheses about what the expressions "Felicia" and "___ is tall" mean. These hypotheses are then tested against further linguistic data and revised if necessary. Learning that George has no sisters, for example, would force me to revise my first version of the Felicia truth rule.

This is where the language of thought is required, Fodor argues. Learning a public language such as English, even if it is your first language, requires you to formulate, test, and revise hypotheses about the truth rules governing individual words. These hypotheses have to be formulated in some language. A truth rule is, after all, just a sentence. But which language are truth rules formulated in?

Fodor thinks that it cannot be the language being learned. You cannot use the language that you are learning to learn that language. That would be pulling yourself up by your own bootstraps. And because Fodor takes his account to apply to children learning their first language no less than to people learning a

second language, the language cannot be any sort of public language. It can only be the language of thought, as described in Chapter 4 .

But how is the language of thought itself learned? The short answer is that it cannot be learned. Because it is a language, the process of learning it would have to involve formulating hypotheses about truth rules (among other things). But Fodor is adamant that those hypotheses can only be formulated in the language of thought. So — you would need to possess a language of thought in order to learn it. He draws the inevitable conclusion, which is that language of thought must be innate. We are all born with a fully developed language of thought, and it is this that allows us to learn the natural language of the community that we are born into.

Fodor ' s argument leads, therefore, to a form of *innatism* or *nativism* about language because he thinks that the very possibility of language learning requires an innate language of thought. There are many other versions of innatism popular among linguists and cognitive scientists. All are based, in one form or another, on arguments to the effect that young children are simply not exposed to enough information, or even the right kind of information, to allow them to learn a language.

These arguments are collectively known as *poverty-of-the-stimulus arguments* . The most famous proponent is the linguist Noam Chomsky. Chomsky ' s view has gone through many iterations, but the basic idea has remained constant.

Chomsky and his followers typically emphasize the following features of the young child ' s learning environment.

- Children are not positively rewarded for language learning.
- Children are typically only exposed to positive information (i.e., they are not told what counts as ungrammatical but only given examples of

grammatical utterances).

- The date from which each child learns is highly idiosyncratic.
- Much of the speech that children hear is actually ungrammatical but is not flagged as such.
- No child encounters more than a tiny fraction of the linguistic information that fixes the grammatical structure of the language.

It would be hard to dispute any of these claims about the learning environment for young children. But do they provide support for some version of innatism?

The best response to an impossibility argument is to give examples of exactly the sort of things that are being claimed to be impossible. In this case, these would be examples of how significant segments of language can be learned with little or no innate knowledge. We will be looking at examples in the remainder of this chapter.

10.3 Language Learning in Neural Networks

Fodor and Chomsky both have a rule-focused approach to linguistic understanding. This section introduces a very different approach based on modeling linguistic abilities in artificial neural networks.

Chapter 5 revealed fundamental differences between artificial neural networks and the sort of computational systems to which the physical symbol system hypothesis applies. In particular:

- Representation in neural networks is distributed across the units and weights, whereas representations in physical symbol systems are encoded in discrete symbol structures.

- There are no clear distinctions in neural networks, either between information storage and information processing or between rules and representations.
- Neural networks are capable of sophisticated forms of learning. This makes them very suitable for modeling how cognitive abilities are acquired and how they evolve.

The second feature, in particular, suggests that neural network models of language are going to be very different from the rules-based approach just discussed. If neural networks do not admit a clear distinction between rules and representations, then they cannot explicitly represent truth rules (or any other type of rules). Despite this, neural networks have been strikingly successful at modeling how languages are learned.

Developmental psychologists and psycholinguists have carefully studied patterns in how children learn languages. They have discovered that in many aspects of language acquisition, children display a very typical trajectory. So, for example, children make very similar types of errors at similar stages in learning particular grammatical constructions. Neural network researchers have explored the extent to which their models can reproduce these characteristic patterns. They have found some striking analogies between how children learn and how neural networks learn. Let ' s look at them.

The Challenge of Tense Learning

One formidable problem confronting children learning a language such as English is that it has both regular and irregular verbs. Some verbs behave in very predictable ways. So, for example, their past tenses are formed according to straightforward rules. The verb " to bat " is a regular verb. In the present tense, we have " I bat. " In the past tense, this becomes " I batted. " There is a

very simple rule here. For regular verbs, we form the past tense by adding the suffix " -ed " to the stem of the verb. The stem of " to bat " is " batt-. " For regular verbs, then, all that one needs to know in order to be able to put them in the past tense is their stem.

Contrast regular verbs with irregular verbs. We have " I give " in the present tense. This becomes " I gave " in the past tense — not " I gived, " as the simple rule might suggest. Likewise for " I take, " which becomes " I took. " Irregular verbs, by their very nature, are not easily summarized by simple rules. It is true that there are observable regularities in how the past tenses of irregular verbs are formed. So, for example, we see that both " I ring " and " I sing " have similar past tenses (" I rang " and " I sang "). It would be unwise, however, to take this as a general rule for verbs ending in " -ing. " The past tense of " I bring " is not " I brang. " Anyone who has ever learned English as a second language will know that the corpus of irregular verbs is full of " false friends " such as these.

Yet somehow, more or less all young children in the English-speaking world manage to find their way through this minefield. How do they do it?

Researchers such as psychologist Stan Kuczaj have studied the grammaticality judgments that children make about sentences involving past-tense verbs in order to examine how their understanding of the past tense develops. The test sentences included both correct past-tense forms (such as " brought " and " gave ") and incorrect ones (such as " brang " and " gived "). The incorrect ones were typically constructed either by treating irregular verbs as if they were regular (as in " gived ") or by exploiting " false friends " (as in " brang "). Looking at patterns of grammaticality judgments across populations of children aged from three to eleven has led researchers to hypothesize that children go through three distinct stages in learning the past tense.

In the first stage, young language learners employ a small number of very common words in the past tense (such as *got* , *gave* , *went* , and *was*). Most of these verbs are irregular, and the standard assumption is that children learn these past tenses by rote. Children at this stage are not capable of generalizing from the words that they have learned. As a consequence, they tend not to make too many mistakes. They can ' t do much, but what they do, they do well.

In the second stage, children use a much greater number of verbs in the past tense, some of which are irregular but most of which employ the regular past-tense ending of " -ed " added to the root of the verb. During this stage, they can generate a past tense for an invented word (such as " rick ") by adding " -ed " to its root. Surprisingly, children at this stage take a step backward. They make mistakes on the past tense of the irregular verbs that they had previously given correctly (saying, for example, " gived " where they had previously said " gave "). These are known as *overregularization* errors.

In the third stage, children cease to make these overregularization errors and regain their earlier performance on the common irregular verbs while at the same time improving their command of regular verbs. Table 10.1 shows the basic trajectory.

Table 10.1 The stages of past-tense learning according to verb type

	STAGE 1	STAGE 2	STAGE 3
Early verbs	Correct	Overregularization errors	Correct
Regular verbs		Correct	Correct
Irregular verbs		Overregularization errors	Improvement with time

This pattern of performance is certainly consistent with Fodor ' s rule-governed conception of language learning. Perhaps, for example, what happens in the second stage is that children make a general hypothesis to the effect that all verbs can be put in the past tense by adding the suffix " -ed " to the root. This hypothesis overrides the irregular past-tense forms learned earlier by rote

and produces the documented regularization errors. In the transition to the third stage, the general hypothesis is refined as children learn that there are verbs to which it does not apply and, correspondingly, begin to learn the specific rules associated with each of these irregular verbs.

Cognitive scientists Steven Pinker and Alan Prince have, in fact, proposed a model of understanding of the English past tense along these lines. Their model has two components and, correspondingly, two information-processing routes. These are illustrated in Figure 10.1 .

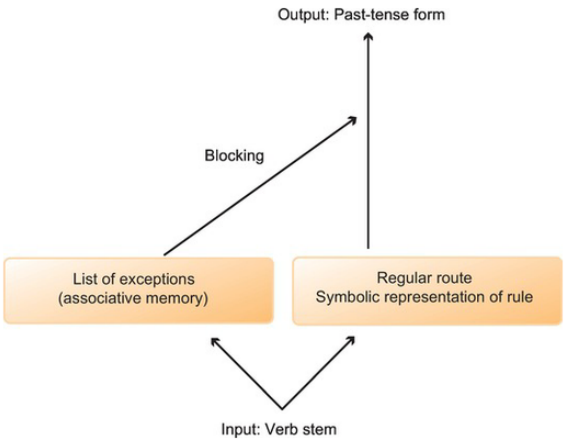


Figure 10.1 The dual-route model of past-tense learning in English proposed by Steven Pinker and Alan Prince.

(From Prince and Pinker 1988.)

One route goes via a symbolic representation of the rule that the past tense is formed by adding " -ed " to the stem of the verb. The symbolic component is not sensitive to the particular phonological form of the verb. It does not recruit information that, for example, the present tense of the verb ends in " -ing. " It simply applies the rule to whatever input it gets.

The second route, in contrast, goes via an associative memory system that is sensitive to the phonological form of the verb stem. It is responsible for storing exceptions to the general rule. It classifies and generalizes these

exceptions in terms of their phonological similarity. One would expect this mechanism to pick up very quickly on the similarity, for example, between *sing* and *ring* .

The two routes compete with each other. The default setting is the symbolic route. The system ′ s " working hypothesis " is that it is dealing with a verb where the past tense is formed by adding " -ed " to the stem. But this default setting can be overridden by a strong enough signal coming from the associative memory system that keeps track of exceptions. What makes signals from the override system strong is that they have been suitably reinforced through experience. If I have had plenty of exposure to the " sing – sang " and " ring – rang " pairs, then this will strengthen the signal for " bring – brang. " But the more exposure I have to the " bring – brought " pair, the weaker the signal for " bring – brang. " Gradually, as I become increasingly exposed to different irregular forms, the signals that are reinforced end up being generally correct.

The model proposed by Pinker and Prince is certainly compatible with the general trajectory of how children learn the English past tense. It is also supported by the general considerations we looked at earlier. But should we accept it (or some other rule-based model like it)?

Exercise 10.2 Explain how this two-component model of past-tense understanding is compatible with the stages identified earlier in young children ′ s learning of the past tense in English.

This is where artificial neural networks come back into the picture because researchers in neural network design have designed networks that reproduce the characteristic pattern of errors in past-tense acquisition without

having programmed into them any explicit rules about how to form the past tense of verbs, whether regular or irregular.

Connectionist Models of Tense Learning

The pioneering network in this area was designed by David Rumelhart and Jay McClelland and appeared in their 1986 collection of papers on parallel distributed processing. A relatively simple network, without any hidden units (and hence not requiring a backpropagation learning algorithm), it nonetheless succeeded in reproducing significant aspects of the learning profile of young children.

As illustrated in Figure 10.2 , there are really three different networks here. The first network takes as input a sequence of phonemes corresponding to the root form of a verb. Phonemes are what linguists take to be the most basic meaningful constituents of words. An example is the phoneme / *n* /, which is the final sound in the words *tin* and *sin* . The first network translates this sequence of phonemes into a representational format that will allow the network to detect relevant similarities between it and other verb roots — as well as between the root forms and the correct past-tense forms.

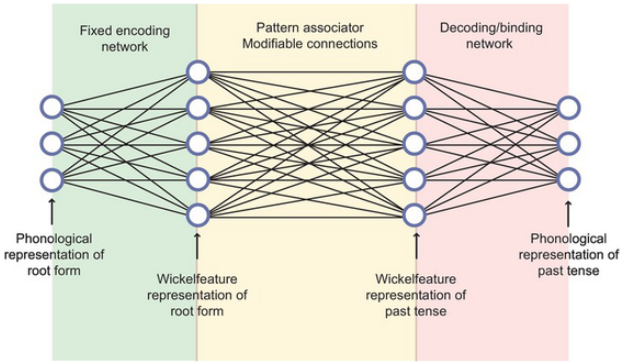


Figure 10.2 Rumelhart and McClelland ′ s model of past-tense acquisition.

(Adapted from Rumelhart et al. 1989 .)

This representational format exploits an ingenious device that Rumelhart and McClelland call *Wickelfeatures* (after the cognitive psychologist Wayne Wickelgren, whose ideas they adapted). A Wickelfeature representation codes phonetic information about individual phonemes within a word and their context. The aim is to represent verb stems in a way that can capture similarities in how they sound (and hence better represent the sort of stimuli to which young children are exposed).

The first network (the network converting phonological representations into Wickelfeature representations) is fixed. It does not change or learn in any way. The learning proper takes place in the second network. This network has no hidden units. It is a simple *pattern-associator* mechanism. It associates input patterns with output patterns. The output patterns are also Wickelfeature representations of words, which are then decoded by the third network. This third network essentially reverses the work done by the first network. It translates the Wickelfeature representations back into sequences of phonemes.

The network was initially trained on ten high-frequency verbs. The aim here was to simulate the first stage in past-tense acquisition. And then it was trained on a much larger training set of 410 medium-frequency verbs (of which 80 percent were regular).

The learning algorithm used by the network is the perceptron-convergence rule that we studied back in Section 5.2 . At the end of the training, the network was almost errorless on the 420 training verbs and generalized quite successfully to a further set of 86 low-frequency verbs that it had not previously encountered (although, as one might expect, the network performed better on novel regular verbs than on novel irregular verbs).

The Rumelhart and McClelland network reproduced the overregularization phenomenon. This is shown in Figure 10.3 , which maps the network ′ s relative success on regular and irregular verbs. As the graph shows,

the network starts out rapidly learning both the regular and the irregular past-tense forms. And then there is a sharp fall in performance on irregular verbs after the eleventh training cycle while at the same time, the degree of success on regular verbs continues to increase.

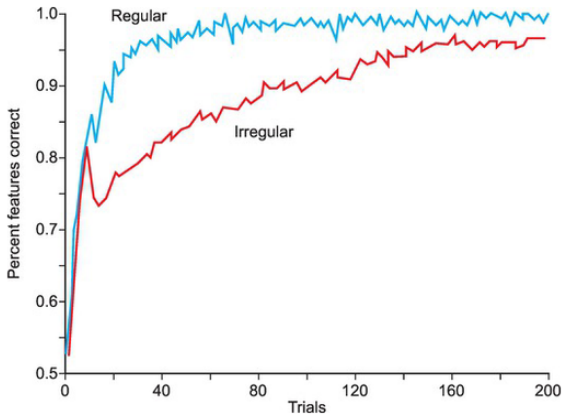


Figure 10.3 Performance data for Rumelhart and McClelland ' s model of past-tense learning. The graph shows the success rates for both regular and irregular verbs. The line for irregular verbs clearly indicates the overregularization phenomenon.

(Adapted from Rumelhart et al. 1986 .)

In other words, while the network ' s " understanding " of irregular verbs is " catching up " with its performance on regular verbs, it is still making characteristic errors. It is treating irregular verbs as if they were regular. So, the network seems to be doing exactly what young children do when they shift from the correct " gave " to the incorrect " gived " as the past tense of " give. "

Exercise 10.3 Explain in your own words why it is significant that the Rumelhart and McClelland network produces the overregularization phenomenon.

Although these results are striking, there were some methodological problems with the design of their study. In particular, as was pointed out in an early critique by Pinker and Prince, the overregularization effect seems to be built into the network. This is because the training set is so dramatically expanded after the tenth cycle. And because the expanded training set is predominantly made up of regular verbs, it has seemed to many that something like the overregularization phenomenon is inevitable.

Nonetheless, a series of further studies achieved similar results to Rumelhart and McClelland with fewer question-begging assumptions. Kim Plunkett and Virginia Marchman, for example, have produced a network with one layer of hidden units that generates a close match with the learning patterns of young children. The network is illustrated in Figure 10.4 .

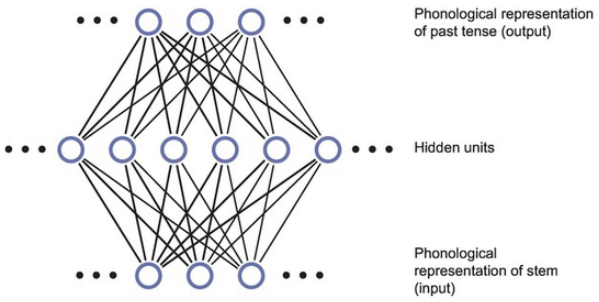


Figure 10.4 The network developed by Plunkett and Marchman to model children ′ s learning of the past tense. The network has a layer of thirty hidden units and is trained using the backpropagation learning algorithm.

(Adapted from Plunkett and Marchman 1993 .)

The Plunkett and Marchman network is, in many ways, a more typical neural network. Whereas the Rumelhart – McClelland network is a simple pattern associator using the perceptron-convergence learning rule, the Plunkett – Marchman model has hidden units. Their model has twenty input units and twenty output units. Between them is a single hidden unit layer with thirty units. The network uses the backpropagation learning algorithm. One advantage of this is that it removes the need to translate the initial phonological representation into Wickelfeatures.

Unlike the McClelland and Rumelhart model, the first stage of the training schedule was on twenty verbs, half regular and half irregular. After that, the vocabulary size was gradually increased. There was no sudden increase — and hence no " predisposition " toward regularization errors. The percentage of regular verbs in the total vocabulary was 90 percent, which more or less matches the relative frequency of regular verbs in English. And yet the network did indeed display the characteristic trajectory, including the regularization errors characteristic of stage 2 learning in children. Plunkett and Marchman correctly guessed that the simple presence in the training set of both regular and irregular verbs would be enough to generate regularization errors during the second stage of training.

It is interesting to compare the learning profile of the Plunkett and Marchman network with the detailed profile of the learning pattern of a child studied by psychologist Gary Marcus. The graph in Figure 10.5 compares the percentage of correctly produced irregular past tenses in the Plunkett and Marchman simulation and in a child whose past-tense acquisition was studied by Marcus and colleagues. As the graph shows, the percentage of correctly produced irregular past tenses drops in both the network and the child as the vocabulary size increases. This seems to correspond to the second of the three

stages identified earlier and seems to be correlated with the predominance of overregularization errors.

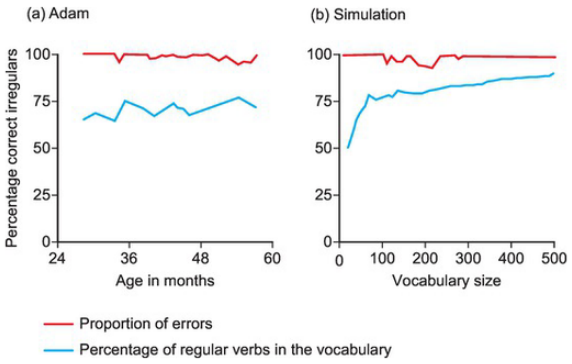


Figure 10.5 A comparison of the errors made by Adam, a child studied by psychologist Gary Marcus, and the Plunkett – Marchman neural network model of tense learning. Unlike the Rumelhart – McClelland model, this model uses hidden units and learns by backpropagation.

(Adapted from McLeod et al. 1998 .)

Certainly, there are huge differences between children learning languages and artificial neural networks learning to correlate verb stems with the correct versions of the past tense. And even when taken on their own terms, neural network models of language acquisition are deeply controversial, not least because of concerns about the biological plausibility of neural networks. But even with these caveats, using artificial neural networks to model cognitive tasks offers a way of putting assumptions about how the mind works to the test — the assumption, for example, that the process of learning a language is a process of forming and evaluating hypotheses about linguistic rules.

The aim of neural network modeling is not to provide a model that faithfully reflects every aspect of neural functioning but, rather, to explore alternatives to dominant conceptions of how the mind works. If, for example, we can devise artificial neural networks that reproduce certain aspects of the

typical trajectory of language learning without having encoded into them explicit representations of linguistic rules, then, at the very least, that suggests that we cannot automatically assume that language learning is a matter of explicitly forming and testing hypotheses about linguistic rules. We should look at artificial neural networks not as attempts to reproduce the mechanics of cognition faithfully but, rather, as tools for opening up new ways of thinking about how information processing might work .

10.4 Bayesian Language Learning

This final section turns to another alternative to explicit, rule-based models of language learning, one applying the Bayesian framework outlined in Chapter 7 . As with the connectionist approach just considered, Bayesian models of language acquisition aim to show how the complexities of language can actually be learned without needing to postulate an innate language of thought or a universal language faculty.

According to Bayesians, language learning takes place through sensitivity to statistical regularities in heard speech, interpreted in terms of the different elements of Bayes ′ s rule — prior probabilities and likelihoods. The basic idea is that young children learning their first language (and adults learning a subsequent language) proceed by updating their probabilities according to Bayes ′ s rule — or, more accurately, that language learning can be modeled as a processing of updating probabilities according to Bayes ′ s rule.

A quick refresher on Bayes ′ s rule is in order. Bayes ′ s rule assumes that we have some evidence. In this case, the evidence is heard speech. It also assumes a hypothesis about how to interpret the evidence. The end result of applying Bayes ′ s rule is to derive what is called a *posterior probability* . The

posterior probability is the probability of the hypothesis in the light of the evidence (or more technically, the probability of the hypothesis conditional on the evidence). It is called *posterior* because it comes *after* considering the evidence.

To apply Bayes ' s rule, we need to have assigned a probability to the *likelihood* of the evidence conditional on the hypothesis. That is, we need to have a view about how likely it is that we would see the evidence that we see if the hypothesis were true. We also need to have assigned a *prior* probability to the hypothesis. Our *prior* is the probability that we assign to the hypothesis *before* considering the evidence.

$$\textbf{With all that in mind, Bayes's rule says that the posterior probability of the hypothesis} = \frac{\textbf{Likelihood of the evidence} \times \textbf{Prior probability of the hypothesis}}{\textbf{Probability of the evidence}}.$$

Bayes ' s rule is a straightforward consequence of the definition of conditional probability.

> **Exercise 10.4** Go back to Chapter 7 and review the discussion of Bayes ' s rule.

Why might one think that Bayes ' s rule would be a useful tool for thinking about language learning? A good starting point is studies showing that young children seem to be very sensitive to statistical regularities in heard speech and, indeed, that adults can use statistical regularities to detect phrase structure.

Probabilities in Word and Phrase Segmentation

A basic challenge in making sense of speech is what is called *word segmentation* . Speech is a continuous stream of sound. In order to make sense of it, the continuous stream of sound needs to be segmented into individual words. This is an essential preliminary to decoding syntactic/grammatical structure and assigning meanings to words and sentences. And in fact, it emerges relatively early in infant development. Developmental linguists hold that word segmentation starts to emerge when infants are around eight months old.

From a phonetic point of view, individual syllables are the natural breaks in the stream of speech. And they are also the building blocks of words. But how does an eight-month-old infant figure out which combinations of syllables make words and which ones do not?

An influential model in developmental linguistics appeals to *transitional probabilities* . Let ′ s take three syllables: / *mo* /, / *ma* /, and / *pa* /. The transitional probability between any two of them is the probability that the second will follow the first. In the language of probability, it is the probability of the second, conditional on the first. So, for example, the transitional probability of / *mo* / and / *ma* / is the probability that / *ma* / follows / *mo* /, whereas the transitional probability of / *mo* / and / *pa* / is the probability that / *pa* / follows / *mo* /.

The basic idea behind applying transitional probabilities to word segmentation is that high transitional probabilities will tend to indicate syllables occurring within a word, whereas low transitional probabilities will tend to occur across the boundaries of words. This makes good sense because there is much greater scope for variation between a syllable at the end of one word and a syllable at the beginning of the next than there is between two syllables occurring within a word.

But where do the probabilities come from? The answer is that they come from frequencies. Young infants have a lot of time to listen to the speech of those around them, sometimes directed at them but most often not. And when they babble, they imitate and reproduce the sounds that they hear. As shown in the 1990s by Jenny Saffran, Richard Aslin, and Elissa Newport, infants, like all higher animals, are exquisitely sensitive to the frequency of correlations, and they can exploit this sensitivity to parse streams of sound into words.

In subsequent work, Elissa Newport and Susan Thompson looked at how transitional probabilities could be a factor in the much more complicated language-learning problem of figuring out how to identify phrases. The basic idea is the same, except that now it is being applied to whole words instead of to individual syllables. The transitional probabilities between two words will typically be higher when those words occur inside a phrase than when they occur across the boundaries of a phrase. So, for example, take any noun phrase —" the camera, " for example. The probability of the word *camera* coming after the word *the* will typically be much higher than the transitional probability of *the* coming after *camera* . This reflects the fact that nouns are typically used with articles (either definite, such as *the* , or indefinite, such as *a*) within a phrase, whereas the occurrence of *the* after *camera* would typically signal the start of a new phrase or even a new sentence.

This work shows the importance of sensitivity to probability and frequencies in language learning. But it is not yet Bayesian reasoning, properly speaking. For that, we need to turn to different examples.

Understanding Pronouns

Linguists have devoted much time and energy to studying pronouns, such as *he* , *she* , *it* , and *one* . They have been particularly intrigued by the phenomenon

typically called *pronominal anaphora* (and in fact, this is a case study often used in poverty-of-the-stimulus arguments). A pronoun is used anaphorically when it picks up on the reference of a term earlier in the sentence. Here is a simple example (or rather, two examples): " Susan ate an ice cream cone, and she enjoyed it. " In this sentence, the pronoun *she* refers to Susan, and *it* refers to the ice cream cone that Susan ate.

It is not difficult to figure out what the two pronouns refer to in " Susan ate an ice cream cone, and she enjoyed it. " Once you ' ve noticed that *she* is a feminine pronoun, everything falls into place very quickly. But many examples of pronominal anaphora are not so straightforward. Consider these two, for example:

1. Tommy beat Mike, and he didn ' t like it.

2. Susan has a black cat, and I want one.

There is no problem interpreting the pronoun *it* in (1). It refers to the fact that Tommy beat Mike. But the second pronoun, *he* , can be interpreted in two different ways. It could refer either to Tommy or to Mike. So, as written, (1) is ambiguous between two possible references for *it* .

Sentence (2) is ambiguous in a different way. What is it that I want? I don ' t want Susan ' s exact cat. But do I specifically want a black cat, like Susan ' s, or am I just generally cat-deprived, so any old cat will do? You can think about it like this. Sentence (2) is ambiguous between two possible membership classes for the pronoun *one* . The pronoun *one* could refer to a member of the class of cats or to a member of the more specific class of black cats.

Typically, though, hearers manage to resolve both ambiguities. How do they do it? This problem has been tackled from a Bayesian perspective. Here is another sentence with anaphoric ambiguity:

3. I ′ ll play with this red ball, and you can play with that one.

The pronoun *one* is ambiguous in (3) in exactly the same way as it was in (2). Are you being instructed to play with a red ball specifically, or with any old ball, which might or might not be red? The sentence still seems ambiguous, even if there is only one other ball in the vicinity and that ball happens to be red. Even if you start playing with the red ball, that in itself doesn ′ t settle whether what I meant was " You can play with that ball " or " You can play with that red ball. "

It turns out that this exact example was proposed by Jeffrey Lidz and collaborators as evidence that infants are born with an innate knowledge of syntactic structure (in an article provocatively titled " What Children Know about Syntax but Could Not Have Learnt "). According to Lidz, it is obvious that the second reading is correct (with the relevant membership class being red balls rather than balls in general). But, he thinks, no evidence available to language users could possibly show that to be the intended reference, for the simple reason that all red balls are balls. This means that the evidence will always confirm both hypotheses. Yet, young children are able to disambiguate this type of sentence from a very early age, and so, he concludes, they must have used innate knowledge to do so.

Responding to Lidz et al., Terry Regier and Susan Gahl showed how Bayesian reasoning could solve the problem. They accept the basic argument that there can be no *direct* evidence to support the red ball hypothesis, but they think that children might have other, *indirect* tools at their disposal. In particular, they point out that, from a Bayesian perspective, we need to look at the likelihoods. In other words, we need to consider how probable the evidence is, given the hypothesis.

Imagine that you hear someone say to you, " I ' ll play with this red ball, and you can play with that one. " You ' ve heard that same sentence a few times before, and every time there has been a red ball there for you to play with. The only evidence you have, therefore, is that the ball you have to play with is red, and as we ' ve said, that is perfectly compatible with the phrase " that one " referring back to the general class of balls rather than to the specific class of red balls.

But you need to think about how likely the evidence is, given the hypothesis. In other words, you need to ask yourself which of these two scenarios is more probable:

a. The balls in the room are always red (*evidence*), if what he means is that I should play with that red ball (*hypothesis*).

b. The balls in the room are always red (*evidence*), if what he means is that I should play with any old ball, red or not (*hypothesis*).

Regier and Gahl make the plausible claim that from a Bayesian perspective, (a) is more probable than (b). This is because it fits the evidence better. If the instruction is to play with any old ball, irrespective of color, then you would expect to see some nonred balls in the room at least some of the time.

Here ' s the analogy they give. Suppose you are trying to decide whether all animals bark (*hypothesis 1*) or whether only dogs bark (*hypothesis 2*). Your evidence is that you have heard lots of dogs barking, but you have never heard anything that is not a dog bark. All dogs are animals, of course, and so your evidence is consistent with both hypotheses. But still, you would be wise to opt for hypothesis 2 because if all animals barked, then you would expect to have run into some barking animals that are not dogs.

One of the key claims of Bayesian approaches to language learning, therefore, is that cognitive scientists have been too quick to make claims about

aspects of language being impossible to learn. Children and adults can learn a lot from thinking about Bayesian likelihoods, as the example of pronominal anaphora shows .

Learning Linguistic Categories

A related problem much studied by linguists and cognitive scientists is how children come to learn linguistic categories and thus distinguish between different types of nouns. This can be particularly challenging when the categories have overlapping membership. To continue with canines, all Dalmatians are dogs, and all dogs are animals. So, we have three linguistic categories:

Dalmation (subordinate category)

Dog (basic category)

Animal (superordinate category)

Imagine a child trying to learn the word *dog* . The family dog happens to be Spotty the Dalmatian, and the child learns to associate the word *dog* with Spotty. Because she knows that Spotty ′ s name is " Spotty, " she can quickly figure out that " dog " is not another name for Spotty. But how does she figure out that " dog " refers to all dogs, not just to Dalmatians — and not to all animals either?

This is another area where cognitive scientists have appealed to innate knowledge, suggesting that children are born with a bias in favor of interpreting category words as basic-level categories, corresponding, for example, to biological species. As before, this has been supported by a general argument that children could not possibly learn to discriminate between subordinate, basic, and superordinate categories simply on the basis of

experience. This argument has been challenged by Fei Xu and Joshua Tenenbaum, among others.

Xu and Tenenbaum ran a set of experiments on category learning in adults to try to uncover tools that young children might be using to learn category words. The experimenters used photographs of objects that fell into subordinate – basic – superordinate taxonomies. One of these taxonomies was, in fact, the Dalmatian – dog – animal taxonomy. They were asked to figure out the meaning of a new, non-English word (e.g., " blick "), based on a small set of labeled samples. This part of the experiment was intended to map the experience of a typical language-learning child.

At the same time, in an effort to understand the tools language users might be using, Xu and Tenenbaum asked their subjects to make pairwise similarity judgments (on a scale of 1 to 9) between the forty-five different objects in the photographs, instructing them to focus on the aspects of the objects that they had used in interpreting the new words. They then used these similarity judgments to draw a hierarchical cluster map of perceived similarities. An example is illustrated in Figure 10.6 . Distance on the map corresponds to perceived similarity. Each node in the tree represents a cluster of stimuli more similar to each other on average than to objects in the nearest cluster.

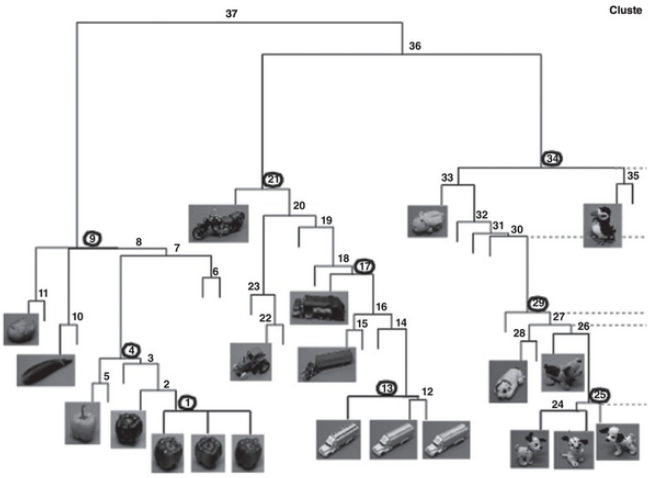


Figure 10.6 A hierarchical cluster of similarity judgments, with nodes corresponding to clusters of stimuli more similar on average to each other than to objects in the nearest cluster.

(From Xu and Tenenbaum 2007 .)

Xu and Tenenbaum then developed a Bayesian model, using this cluster analysis as the hypothesis space. Each similarity cluster corresponded to a single hypothesis about the possible extension of a word such as " blick, " on the principle that one can ' t view objects as falling under a single category word without seeing them as similar in certain respects. This means that any set of objects viewed as similar is a candidate for the extension of a given word.

But how, given a range of hypotheses (corresponding to clusters of objects perceived as similar), is one going to be decided upon? Xu and Tenenbaum propose a Bayesian decision process, exploiting the key Bayesian concepts of *prior probability* and *likelihood* . They used the cluster space to derive priors and likelihoods.

They derived the priors on the assumption that the more distinctive a similarity cluster is, the more likely it is to have its own name. This makes good sense. Languages contain names to pick out salient groups of things, after all, and the more distinctive a similarity cluster is, the more it will stand out (be

salient) — and the easier it will be for language users to pick out members of the group.

The likelihoods for the model were arrived at by applying what they called the *size principle* . This is, in effect, the same principle that we looked at in the context of red balls and barking dogs. It basically says that you should interpret the hypothesis being confirmed as narrowly as possible.

So, the hypothesis that " blick " refers to Dalmatians is far more likely than the hypothesis that it refers to dogs, if you have heard it applied to three Dalmatians in a row — and much more likely still than the hypothesis that " blick " refers to animals in general. Remember that the likelihood measures the probability that the evidence would occur, given the hypothesis. And what the size principle says is that a narrow hypothesis (relative to the evidence) is more likely to generate that evidence than a broader one.

The key finding from Xu and Tenenbaum ' s experiments on adults was that their category-membership judgments were accurately predicted by a Bayesian model using the priors and likelihoods just described — that is, with priors fixed by the similarity space and likelihoods by the size principle. Moreover, when they went on to test three- to four-year-old children on a similar task, they found that the children ' s categorization judgments closely mapped those of the adults, which is certainly consistent with the hypothesis that their linguistic behavior could be captured by a Bayesian model (assuming that their similarity judgments mapped those of adults).

The Xu and Tenenbaum experiments provide another illustration of how Bayesian statistical inference might be deployed by language learners to solve problems that many cognitive scientists have taken to be unsolvable by any type of learning .

Summary

This chapter has explored different ways of modeling language mastery and language learning. We started out with a very general argument, attributable to Jerry Fodor, that language learning must be rule based, and we saw how that line of thinking led him to a form of innatism about language. Chomsky reaches a similar view about language learning, from a rather different starting point. We then turned to two very different approaches. With respect to the specific problem of how children learn the past tense of English verbs, we saw how connectionist models of tense learning offer an alternative to the idea that grammar is learned by internalizing explicitly represented grammatical rules. The final section looked at Bayesian models of language learning and showed how a Bayesian approach can illuminate word segmentation, pronominal anaphora, and learning of hierarchically organized category words.

Checklists

Language and Rules

1. Language is a paradigmatically rule-governed activity.

2. The default hypothesis in thinking about language learning is that it is a matter of learning the rules that govern the meanings of words and how they combine into meaningful units.

3. Fodor has built on the default hypothesis to argue that learning a language requires learning *truth rules* , which must be stated in the language of thought.

4. According to Fodor, the language of thought cannot itself be learned and so must be innate.

5. Noam Chomsky reached a similar nativist/innatist conclusion based on poverty-of-the-stimulus arguments.

6. One way to challenge such arguments is to construct models that simulate the trajectory of human language learning without explicitly representing any rules.

Modeling the Acquisition of the English Past Tense

1. Children learning the English past tense go through three easily identifiable stages:

Stage 1: They employ a small number of verbs with (mainly irregular) past tenses.

Stage 2: They employ many more verbs, tending to construct the past tense through the standard stem + " -ed " construction (producing errors on verbs they had formerly got right).

Stage 3: They learn more verbs and correct their overregularization errors.

2. This pattern of past-tense acquisition can be accommodated by a symbolic model.

3. But connectionist models of past-tense acquisition have been developed that display a similar trajectory without having any rules explicitly coded in them.

Bayesian Language Learning

1. The basic Bayesian idea is that language learning can be modeled as a process of updating probabilities according to Bayes ′ s rule.

2. The Bayesian approach is supported by studies showing that both children and adults are highly sensitive to statistical regularities in heard speech.

3. Studies suggest that infants use transitional probabilities for word segmentation (parsing heard speech into words) and that transitional probabilities are also used by adults to map the boundaries of phrases.

4. Pronominal anaphora (identifying the antecedent of a pronoun) has been proposed as a particular challenge for non-innatist approaches to language learning, but the problem seems susceptible to a broadly Bayesian solution.

5. Another traditional problem for empiricist theories of language learning is how children learn hierarchical category words that pick out overlapping sets of objects. Xu and Tenenbaum used a Bayesian model with appropriate priors and likelihoods to capture the linguistic behavior of adults and three- to four-year-old children.

Further Reading

Fodor ′ s discussion of truth rules is in his book *The Language of Thought* (Fodor 1975). Chomsky first raised poverty-of-the-stimulus arguments in his influential review of B. F. Skinner ′ s book *Verbal Behavior* in 1959. See also Chomsky 1968 , 1980a (summarized with comments and critique in Chomsky 1980b) and, for more recent discussion, Berwick et al. 2011 . For a critical

discussion of these arguments, see Pullum and Scholz 2002, and for a general discussion of arguments for nativism, see Cowie 1999.

The second volume of *Parallel Distributed Processing* (McClelland et al. 1986) contains a number of papers applying the theoretical framework of connectionism to different cognitive abilities. Some of these applications are explored further in McLeod et al. 1998 and Plunkett and Elman 1997 . For a more general discussion of modeling within a connectionist framework, see Dawson 2004 . Paul Churchland has been a tireless proponent of the power of connectionist networks; see, for example, the papers in Churchland 2007 for a wide range of applications. See also McClelland et al. 2010 .

Chapter 18 of the original parallel distributed processing (PDP) collection (Rumelhart and McClelland 1986) was the first salvo in what has become a lengthy debate about how to model past-tense learning. Pinker and Prince 1988a made some telling criticisms of Rumelhart and McClelland ′ s model (Pinker and Prince 1988b , reprinted in Cummins and Cummins 2000 , is more condensed). A number of researchers took up Pinker and Prince ′ s challenge — see, for example, Plunkett and Marchman 1993 . The work by Marcus described in the text is presented in Marcus et al. 1992 . For a more recent exchange, see Pinker and Ullman 2002 and the reply in McClelland and Patterson 2002 . Connectionist models have been applied to many different aspects of language. Plaut et al. 2003 describes applications to phonology, morphology, and syntax. Christiansen and Chater 2001 is an interdisciplinary collection of papers in the emerging field of connectionist psycholinguistics. Westermann and Ruh 2012 provides a review of different approaches to past-tense learning, including connectionist approaches. Perhaps the most famous formal result in the theory of language learning is Gold ′ s theorem, which places constraints upon the class of languages that can be learned with purely positive feedback. Gold ′ s theorem is clearly presented in Johnson 2004 .

Doug Rohde and David Plaut have used neural network models to argue that Gold ' s theorem cannot straightforwardly be applied in cognitive science (Rohde and Plaut 1999).

For a helpful and up-to-date overview of Bayesian approaches to language acquisition, see Pearl and Goldwater 2016 . See also Chater and Manning 2006 . Saffran et al. 1996 and Aslin et al. 1998 were pioneering studies of word segmentation in infancy, mainly working with artificial languages. Pelucchi et al. 2009 extended the approach to more realistic linguistic situations. Thompson and Newport 2007 applies transitional probabilities to phrase segmentation. The Bayesian analysis of pronominal anaphora described in the text originates in Regier and Gahl 2004 , who were responding to Lidz et al. 2003 . The Xu and Tannenbaum experiments on category learning are described in Xu and Tenenbaum 2007 . For a recent extension of this approach to one-shot category learning, see Lake et al. 2016 .

Chapter 11

Object Perception and Folk Physics



Overview

11.1 Object Permanence and Physical Reasoning in Infancy

Infant Cognition and the Dishabituation Paradigm

How Should the Dishabituation Experiments Be Interpreted?

11.2 Neural Network Models of Children's Physical Reasoning

Modeling Object Permanence

Modeling the Balance-Beam Problem

11.3 Conclusion: The Question of Levels

Overview

Back in Chapter 5 , we looked at information processing in single-unit networks and then saw how the power of neural networks increases when hidden units are added. In Chapter 10 , we started exploring how neural networks can model cognition. We looked at neural network models of past-

tense learning and saw how their learning trajectory bears striking resemblances to the learning trajectory of human infants. This chapter turns to another application of neural networks. We will see how they can be used to model object perception, particularly what developmental psychologists call *object permanence* .

Many studies have shown that the perceptual universe of human infants is far more complex and sophisticated than traditionally thought. From a very early age, human infants are sensitive to certain basic properties of physical objects. They have definite (and often accurate) expectations about how objects behave and interact. Some of this research is presented in Section 11.1 , where we see how it can very naturally be interpreted in computational terms as involving an explicitly represented and quasi-theoretical body of rules and principles (a *folk physics*).

In Section 11.2 , however, we show how some of the very same data can be accommodated without this type of explicit, symbolic representation. We look at neural network models that share some of the basic behaviors of the infants in the experiments without having any rules or principles explicitly coded into them. This opens the door to a different way of thinking about infants ′ knowledge of the physical world.

Finally, in Section 11.3 , we turn to an issue that is in the background in the earlier sections of this chapter and Chapter 10 . How exactly should we think about the relation between symbolic models (physical symbol systems) and neural network models? Are they in competition with each other? Or are they giving accounts on different levels? We explore these issues through a famous objection to neural network modeling attributable to Jerry Fodor and Zenon Pylyshyn.

11.1 Object Permanence and Physical Reasoning in Infancy

What is it like to be a human infant? Until the last decades of the twentieth century, most developmental psychologists were convinced that the infant experience of the world is fundamentally different from our own. The famous psychologist and philosopher William James (brother of the novelist Henry James) coined the memorable phrase " a blooming, buzzing, confusion " to describe what it is like to be a newborn infant (a *neonate* , in the jargon of developmental psychologists). According to James, neonates inhabit a universe radically unlike our own, composed solely of sensations, with no sense of differentiation between self and objects or between self and other and in which the infant is capable only of reflex actions. It takes a long time for this primitive form of existence to become the familiar world of people and objects and for reflexes to be replaced by proper motor behavior.

The most famous theory within the traditional view was developed by Swiss psychologist Jean Piaget (1896 – 1980). According to Piaget, infants are born with certain innate, reflex-like sensorimotor schemas that allow them to perform very basic acts, such as sucking a nipple. Infants gradually bootstrap these basic schemas into more complex behaviors (what Piaget called *circular reactions*) and gradually come to learn that they inhabit a world containing other objects and other individuals. According to Piaget, infants are born highly egocentric, and it is not until the end of what he called the *sensorimotor stage* (at around two years of age) that they come fully to appreciate the distinctions between self and other and between the body and other physical objects.

In recent years, however, researchers have developed new techniques for studying the cognitive abilities of neonates and older infants. These techniques have led to a radical revision of the traditional view. As a consequence, many developmental psychologists now think that the world of the human infant is much less of a " blooming, buzzing, confusion " than James thought. It is now widely held that even very young infants inhabit a highly structured and orderly perceptual universe. The most famous technique in this area is called the *dishabituation paradigm* , which is a technique for exploring the expectations that infants have about how objects will behave.

Infant Cognition and the Dishabituation Paradigm

The basic idea behind the dishabituation paradigm is that infants look longer at events that they find surprising. So, by measuring the amount of time that infants look at events of different types, experimenters can work out which events the infants find surprising and then use this to work backward to the expectations that the infants had about how those events were going to turn out.

So, if infants look longer at something, then that suggests that it did not turn out the way they expected. This basic idea is applied in practice in a number of ways. One technique is to habituate infants to a given type of event (i.e., presenting the infants with examples until they lose interest) and then to present them with events that differ from the original one in certain specified ways. Looking-time measures can then be used to identify which of the new events capture the infants ' attention, as measured by the amount of time the infants spend looking at them. This allows experimenters to

detect which features of the events the infants find surprising — and hence to work out how the infants expected the events to unfold. This way of identifying " violation of expectations " is called the *dishabituation paradigm* .

The *drawbridge* experiments designed by the developmental psychologist Renée Baillargeon illustrate how the paradigm works and what we can learn from it about the perceptual universe of the human infant. In one set of experiments, Baillargeon habituated her infants (who were all about 4.5 months old) to a screen (the drawbridge) rotating 180 degrees on a table. She was interested in how the infants would react when an object was hidden within the drawbridge ' s range of motion because this would be a way of finding out whether the infants had any expectations about objects they could not directly perceive.

In order to investigate this, Baillargeon contrived a way of concealing the object so that although it could not be seen by the infant, any adult or older child looking at the apparatus could easily work out that it would obstruct the movement of the screen. She then presented infants with two different scenarios. In the first scenario, the screen rotated as it had done before until it got to the place where the obstructing box would be — and then it stopped, exactly as you or I would expect it to. In the second scenario, the screen kept on rotating for the full 180 degrees and hence apparently passed through the obstructing box. The experiments are illustrated in Figure 11.1 .

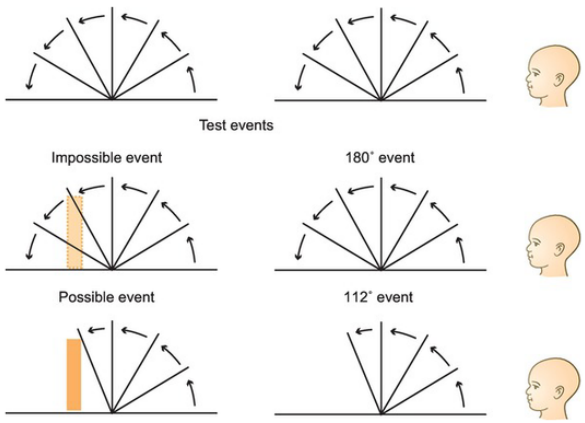


Figure 11.1 Schematic representation of the habituation and test conditions in Baillargeon ′s drawbridge experiments. After habituation to a drawbridge moving normally through 180 degrees, infants were tested both on an impossible event (in which the drawbridge ′s movement would require it to pass through a hidden object) and a normal event (in which the drawbridge halts at the point where it would make contact with the hidden object). Baillargeon found that 4.5-month-old infants reliably looked longer in the impossible condition.

(Adapted from Baillargeon 1987.)

Baillargeon found that the infants looked significantly longer in the second scenario. They were, it seemed, surprised that the screen looked as if it was passing straight through the obstructing box. So, if we assume that infants look longer when their expectations are violated, the experiments show that they do not expect the screen to keep on rotating through the place where the obstructing box would be. Baillargeon concluded that although the infants could not see the obstructing box, in some sense, they nonetheless “ knew ” that the box was there — and that the screen could not pass through it.

This result has direct implications for a long-running debate in developmental psychology. Developmental psychologists have long been concerned with the following question: At what stage, in early childhood or infancy, is it appropriate to ascribe a grasp that objects exist even when not being perceived? This is what is typically called *object permanence* . In the traditional view, derived ultimately from Piaget, object permanence does not appear until relatively late in development, at about eight or nine months. What Baillargeon ′ s drawbridge experiments seem to show, however, is that object permanence emerges much earlier than Piaget (and others) had thought .

But there is more going on here. It is not just that the infants are in some sense aware that the obstructing box is there even though they cannot see it. Their surprise at the second scenario shows that they have expectations about how objects should interact. In fact, Baillargeon ′ s drawbridge experiments, together with other experiments using the same paradigm, have been taken to show that even very young infants have the beginnings of what is sometimes called *folk physics* (or *naive physics*) — that is to say, an understanding of some of the basic principles governing how physical objects behave and how they interact.

Elizabeth Spelke is another pioneer in using dishabituation experiments to study the perceptual universe of human infants. She has used a battery of experiments to argue that from a very young age, infants are able to parse the visual array into spatially extended and bounded individuals. These individuals behave according to certain basic principles of physical reasoning. She thinks that four of these principles are particularly important for understanding the infant ′ s *folk physics* or *intuitive physics* .

First is the *principle of cohesion* , according to which surfaces belong to a single individual if and only if they are in contact. For example, infants do not appear to perceive the boundary between two objects that are stationary and adjacent, even when the objects differ in color, shape, and texture. Figure 11.2 illustrates how sensitivity to the principle of cohesion might be experimentally tested.

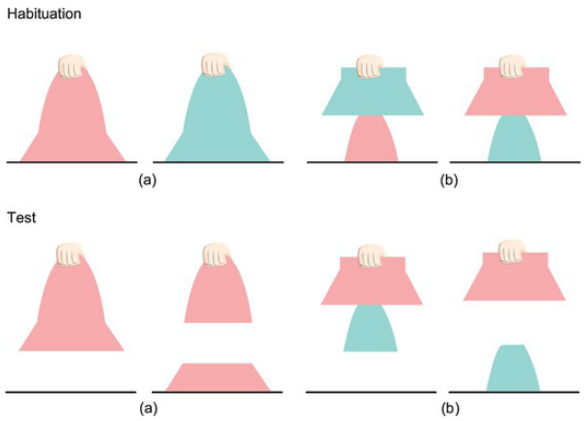


Figure 11.2 Schematic representation of an experiment used to test infants ′ understanding of object boundaries and sensitivity to Spelke ′ s principle of cohesion (that surfaces lie on a single object if they are in contact).

(Adapted from Spelke and Van de Walle 1993.)

Three-month-old infants are habituated to two objects, one more or less naturally shaped and homogeneously colored, and the other a gerrymandered object that looks rather like a lampshade. When the experimenter picks up the objects, they either come apart or rise up cleanly. Infants show more surprise when the object comes apart, even if (as in the case of the lampshade) the object does not have the Gestalt properties of homogeneous color and figural simplicity. The conclusion drawn by Spelke

and other researchers is that the infants perceive even the gerrymandered object as a single individual because its surfaces are in contact.

The principle of cohesion suggests that infants will perceive objects with an occluded center as two distinct individuals because they cannot see any connection between the two parts. And this is indeed what they do — at least when dealing with objects that are stationary. Thus, it seems that infants do not perceive an occluded figure as a single individual, *if the display is static* . After habituation to the occluded figure, they showed no preference for either of the test displays.

On the other hand, however, infants do seem to perceive a center-occluded object as a single individual if the object is in motion (irrespective, by the way, of whether the motion is lateral, vertical, or in depth). According to Spelke, this is because there is another principle at work. According to the *principle of contact* , only surfaces that are in contact can move together. When the principle of cohesion and the principle of contact are taken together, they suggest that because the two parts of the occluded object move together, they must be in contact and hence, in fact, be parts of one individual. This is illustrated in Figure 11.3 .

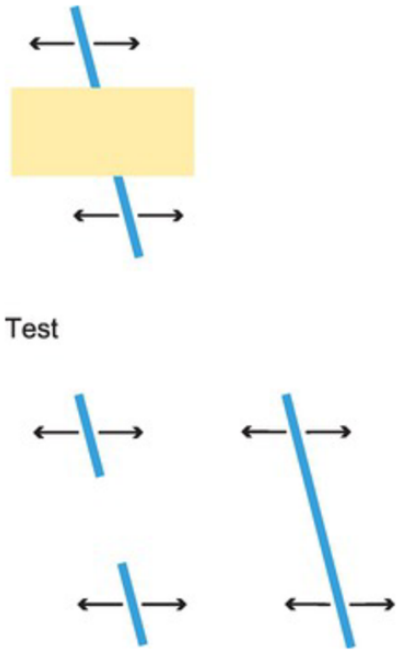


Figure 11.3 Schematic representation of an experiment testing infants ' understanding of the principle of contact (that only surfaces in contact can move together).

(Adapted from Spelke and Van de Walle 1993.)

Exercise 11.1 Explain how an infant who understands the principles of cohesion and contact might respond to the two test situations depicted in Figure 11.3 .

Spelke identifies two further constraints governing how infants parse the visual array. It is generally true of physical objects that every object moves on a single trajectory through space and time, and those paths cannot intersect in a way that would allow more than one object to be in one place at a time. One can test whether infants are perceptually sensitive to these features by investigating whether they are surprised by breaches of what

Spelke calls the *solidity* and *continuity* constraints. The drawbridge experiment that we have just discussed is a good example of reasoning according to the solidity constraint because it shows that infants are sensitive to the impossibility of there being more than one object in a single place at one time. Figure 11.4 is a schematic representation of an experiment to test whether infants parse their visual array in accordance with the continuity and solidity constraints.

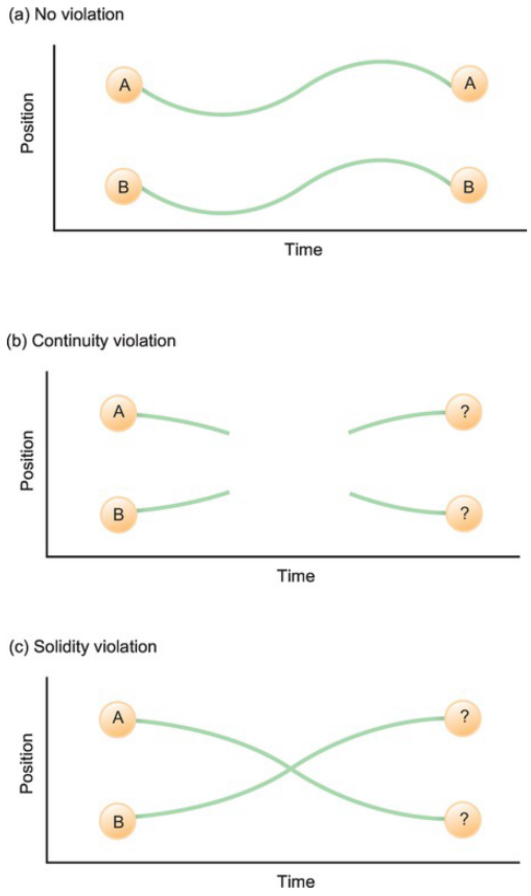


Figure 11.4 Schematic depiction of events that accord with, or violate, the continuity or solidity constraints. Solid lines indicate each object ′s path of motion, expressed as changes in its position over time. Each object traces (a) exactly one connected path through space and time, (b) no connected path through space and time, or (c) a path through space and time intersecting another object ′s path.

(Adapted from Spelke and Van de Walle 1993.)

These are just some of the experiments showing that even very young infants have a surprisingly sophisticated understanding of the physical world. Spelke herself has some definite views about what this understanding consists of. According to Spelke, even very young infants have a theoretical understanding of physical objects and how they behave.

Infants are able to represent principles such as those that we have been discussing — the principles of continuity, solidity, and so on. They can use these principles to make predictions about how objects will behave. They show surprise when those predictions are not met — and lose interest (as measured by looking times) when they are met.

How Should the Dishabituation Experiments Be Interpreted?

What the infants are doing, according to Spelke (and many others), is not fundamentally different in kind from what scientists do. The infants are making inferences about things that they cannot see on the basis of effects that they can see — just as scientists make inferences about, say, subatomic particles on the basis of trails in a cloud chamber. Infants are little scientists, and the perceptual discriminations that they make reflect their abilities to make inferences about the likely behavior of physical objects, inferences that in turn are grounded in a stored and quasi-theoretical body of knowledge about the physical world — what is sometimes called *infant folk physics* .

So, what sort of information processing underlies infant folk physics? The physical symbol system hypothesis gives us a natural way of thinking about how rules might be explicitly represented and applied. The idea here is that the basic principles of infant folk physics (such as the principle of continuity) are symbolically represented. These symbolically represented principles allow the infants to compute the probable behavior and trajectory of the objects in the dishabituation experiments. Infants show surprise when objects do not behave according to the results of the computations.

At the same time, infant folk physics is importantly different from adult folk physics. Developmental psychologists have found, for example, that infants tend to place more weight on spatiotemporal continuity than on featural continuity. For infants, movement information dominates information about features and properties. Their principal criterion for whether or not an object persists over time is that it should maintain a single trajectory, even if its perceptible properties completely change.

This is why, for example, infants who otherwise perceive differences between the color and form of objects still tend not to show surprise when one object disappears behind a screen and another completely different object emerges at the other side of the screen. For adults, on the other hand, featural constancy is often more important. This is elegantly expressed by developmental psychologists Alison Gopnik and Andrew Meltzoff:

> As adults we individuate and reidentify objects by using both place and trajectory information and static-property information. We also use property information to predict and explain appearances and disappearances. If the same large, distinctive white rabbit appears in the box and later in the hat, I assume it ′ s the same rabbit, even if I don ′ t immediately see a path of movement for it. In fact, I infer an often quite complex invisible path for the object. If I see the green scarf turn into a bunch of flowers as it passes through the conjuror ′ s hand while maintaining its trajectory, I assume it is a different object. On the other hand, if an object changes its trajectory, even in a very complex way, while maintaining its properties, I will assume it is still the same object.
>
> (Gopnik and Meltzoff 1997 : 86)

So, there are some important differences between infant folk physics and adult folk physics. The important point, though, is that for Spelke (and indeed, for Gopnik and Meltzoff), both should be understood as theories. Here is how Spelke described her findings in an influential early paper:

> I suggest that the infant ' s mechanism for apprehending objects is a mechanism of thought: an initial *theory* of the physical world whose four principles jointly define an initial *object concept* .
>
> (Spelke 1988 : 181)

It is no easy matter to say what a theory actually is, but as Spelke states, the simplest way of thinking about theories is in terms of laws or principles. Laws and principles can be linguistically expressed. This means that they can easily be represented by physical symbol structures. In this respect, thinking about naive physics as a theory is rather like thinking of grammatical knowledge in terms of rules. In both cases, we have cognitive capacities (knowledge of a theory in the one case, the ability to apply rules in the other) that lend themselves to being modeled in computational terms — as suggested by the physical symbol system hypothesis .

As we saw in the case of grammatical knowledge, however, there are alternatives to this type of computational approach. We can think about it in nonsymbolic ways, exploiting neural network models. Some of the possibilities are sketched out in the next section .

11.2 Neural Network Models of Children

’s Physical Reasoning

Connectionist modelers have explored alternatives to the theoretical model of infant cognitive abilities that we have just looked at. They have tried to show how a neural network can simulate the behavior of human infants in experiments using the dishabituation paradigm without any relevant principles or rules being explicitly coded into it.

One prominent researcher in this area is the psychologist Yuko Munakata, working with a number of collaborators, including the distinguished connectionist modeler Jay McClelland (who, together with David Rumelhart, edited the two-volume *Parallel Distributed Processing* , which gave such a significant impetus to connectionist approaches to cognitive science). Here is how Munakata and her coauthors describe the basic idea behind their approach and how it differs from the theoretical model:

> Because infants seem to behave in accordance with principles at times, there might be some use to describing their behavior in these terms. The danger, we believe, comes in the tendency to accept these descriptions of behavior as mental entities that are explicitly accessed and used in the production of behavior. That is, one could say that infants ’ behavior in a looking-time task accords with a principle of object permanence, in the same way one could say that the motions of the planets accord with Kepler ’ s laws. However, it is a further — and we argue unfounded — step to then conclude that infants actually access and reason with an explicit representation of the principle itself.
>
> (Munakata et al. 1997 : 687)

The connectionist modelers accept that the dishabituation experiments show that human infants are sensitive to (and react in accordance with) certain basic physical principles (such as the principles of solidity and continuity). But they reject how computational theorists interpret this sensitivity.

Computational theorists assume that a cognitive system can only act in accordance with, say, the principle of continuity if that principle is explicitly represented in it in a symbolic form. But according to Munakata and her collaborators, this assumption is wrong — and it can be shown to be wrong by constructing a neural network model that acts in accordance with the principle of continuity even though it does not have that principle symbolically encoded in it. They continue:

> We present an alternative approach that focuses on the adaptive mechanisms that may give rise to behavior and on the processes that may underlie change in these mechanisms. We show that one might characterize these mechanisms as behaving in accordance with particular principles (under certain conditions); however, such characterizations would serve more as a shorthand description of the mechanism ′ s behavior, not as a claim that the mechanisms explicitly consult and reason with these principles.
>
> (Munakata et al. 1997 : 687)

Their alternative proposal is that infants ′ understanding of object permanence is essentially practical. The fact that infants successfully perform object-permanence tasks does indeed show that they know, for example, that objects continue to exist even when they are not being directly perceived. But this knowledge is not *explicitly* stored in the form of

theoretical principles. Rather, it is *implicitly* stored in graded patterns of neural connections that evolve as a function of experience.

In this view, infants ′ expectations about how objects will behave are driven by patterns of neural activation. These patterns vary in strength because of the following factors:

- The number of neurons firing
- The strength and number of the connections between them
- The relations between their individual firing rates

So, the perceptual sensitivity that we see in the dishabituation paradigms results from associative mechanisms of pattern recognition. This is exactly what connectionist networks model so well.

Here is how the process works. When infants observe the " reappearance " of occluded objects, this strengthens the connection between two groups of neurons — between the group of neurons that fire when the object first appears, on the one hand, and the group that fires when it reappears, on the other. As a result, the representations of perceived objects (i.e., the patterns of neural activation that accompany the visual perception of an object) persist longer when the object is occluded. So, according to Munakata et al., the infant ′ s " knowledge " of object permanence should be understood in terms of the persistence of object representations rather than in terms of any explicitly coded principles. This " implicit " understanding of object permanence is the foundation for the theoretical understanding that emerges at a much later stage in development.

Their approach explains well-documented behavioral dissociations in infant development. Infants ′ abilities to act on occluded objects lag a long

way behind their perceptual sensitivity to object permanence. Perceptual sensitivity to object permanence emerges at around four months, but infants succeed in searching for hidden objects only at around eight months. Munakata et al. argue (and their simulations illustrate) that it is possible for a visual object representation to be sufficiently strong to generate expectations about the reappearance of an occluded object while still being too weak to drive searching behavior.

Modeling Object Permanence

One network studied by Munakata et al. simulates a simple object-permanence task where a barrier moves in front of a ball and occludes it for a number of time steps. Figure 11.5 shows the inputs to the network as the barrier moves in front of the ball and then back to its original location. The input units are in two rows. The bottom layer represents the network ' s view of the barrier, and the top layer represents its view of the ball. When the barrier moves in front of the ball, there is no input in the ball layer. When the barrier moves to one side, revealing the previously occluded ball, the ball layer is correspondingly activated again. The network has to learn to represent the ball even when there is no corresponding activation in the input layer — it needs to find a way of representing the ball even when the ball cannot directly be seen.

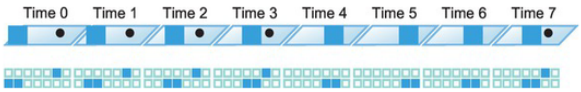


Figure 11.5 A series of inputs to the network as a barrier moves in front of a ball and then back to its original location. At the top is a schematic drawing of an event in the network ' s visual field; below is the corresponding pattern of activation presented to the network ' s input units, with each square representing one unit.

(Adapted from Munakata et al. 1997.)

In order to design a network that can do this, Munakata and her collaborators used a *recurrent network* . These are rather different from the networks we have been considering up to now. Like feedforward and competitive networks, they have hidden units whose weights are modified by algorithmic learning rules. But unlike them, recurrent networks have a feedback loop that transmits activation from the hidden units back to themselves before the learning rule is applied. This feedback loop allows the network to preserve a " memory " of the pattern of activation in the hidden units at the previous stage.

As shown in Figure 11.6 , their recurrent network has two distinctive features. The first is the set of recurrent weights from the hidden layer back to itself. These function as just described — to give the network information about what happened at the previous temporal stage. The second is a set of connections, with corresponding weights, running from the hidden units to the input units. These weighted connections allow the network to send a prediction to the input units as to what the next set of inputs will be. The network ' s learning (which works via the standard

backpropagation rule) is driven by the discrepancy between the actual input and the predicted input.

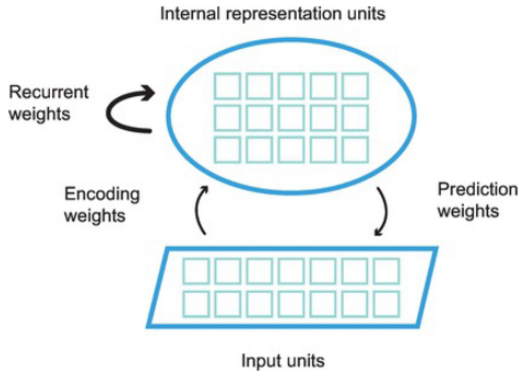


Figure 11.6 Recurrent network for learning to anticipate the future position of objects. The pattern of activation on the internal representation units is determined by the current input and by the previous state of the representation units by means of the encoding weights and the recurrent weights, respectively. The network sends a prediction back to the input units to predict the next state of the input. The stimulus input determines the pattern of activation on the input units, but the difference between the pattern predicted and the stimulus input is the signal that drives learning.

(Adapted from Munakata et al. 1997.)

We can think about the network ' s " understanding " of object permanence in terms of its sensitivity to the ball ' s reappearance from behind the occluder. This sensitivity can in turn be measured in terms of the accuracy of the network ' s " prediction " when the ball does eventually reappear. (An accurate prediction is one where the predicted pattern exactly matches the input pattern.) As training progresses, the network becomes

increasingly proficient at predicting the reappearance of occluded objects over longer and longer periods of occlusion.

What makes this possible is the recurrent connection from the hidden layer back to itself. The activation associated with the " sight " of the ball at a given temporal stage is transmitted to the next stage, even when the ball is not in view. So, for example, at temporal stages 4, 5, and 6 in Figure 11.5 , there is no activation in the input units representing the ball. But once the network ' s training has progressed far enough, the weights will work in such a way that the memory from the earlier stages is strong enough that the network will correctly predict the reappearance of the ball at temporal stage 7.

How exactly does this work? The researchers found that improved sensitivity to object permanence is directly correlated with the hidden units representing the ball showing similar patterns of activation when the ball is visible and when it is occluded. In effect, they claim, the network is learning to maintain a representation of an occluded object. The network ' s " understanding " of object permanence is to be analyzed in terms of its ability to maintain such representations. And this comes in degrees. As further simulations reported in the same paper show, a network can maintain representations sufficiently strong to drive perceptual " expectations " but too weak to drive motor behavior. Sensitivity to object permanence is, they suggest, a graded phenomenon — a function of strengthened connections allowing maintained activation patterns — rather than a theoretical achievement.

Exercise 11.2 Explain and assess the significance of this network model for thinking about the information processing underlying object permanence .

Modeling the Balance-Beam Problem

Here is a second example of how connectionist models can provide alternatives to theory-based accounts of infant cognitive development. This is the balance-beam problem.

Children are shown a balance beam as in Figure 11.7 . The balance beam has a fulcrum and weights at varying distances from the fulcrum. The children are asked whether the beam is in balance, and if not, which side will go down. In different trials the weights are varied, but the children are not given any feedback on whether their answers are correct or not.

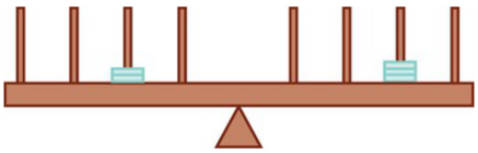


Figure 11.7 A balance beam. Weights can be added at different distances from the fulcrum. Children are asked whether the beam is in balance, and if not, which side will go down.

(From Elman et al. 1996.)

Research by developmental psychologist Bob Siegler has shown that children typically go through a series of stages in tackling the balance-beam problem — rather like young children learning the past tense of English verbs. And as in the past-tense case, these stages can be summarized in

terms of some relatively simple rules. There are four stages and corresponding rules. Siegler identifies these as follows:

> *Stage 1.* Children think that the side with the greatest number of weights will go down, irrespective of how those weights are arranged. If there are equal numbers of weights on both sides, then the beam is judged to be in balance.
>
> *Stage 2.* Now they think that when the weights on each side of the fulcrum are equal, the side on which the weights are farthest away will go down. If this doesn ′ t hold, then children either use the first rule or guess.
>
> *Stage 3.* Now children are able to use the correct rule, understanding that downward force is a function of both weight and the distance from the fulcrum. But they only manage to do this when the two sides differ in respect *either* to weight *or* to distance but not both.
>
> *Stage 4.* It is usually not until adolescence that children acquire a general competence for balance-beam problems — and even then, not all of them do.

The situation here is very similar to the past-tense case. And as in that case, it seems initially plausible to model the child ′ s learning process as a matter of learning a series of rules. But as we ′ ve already seen, there are other ways to think about this type of developmental progression. Artificial neural networks offer an alternative way of looking at the phenomenon, illustrating how the appearance of rule-based learning can emerge from a system that does not exploit any explicit rules.

Jay McClelland and E. Jenkins designed an artificial neural network to model children ' s performance on the balance-beam problem. The network, illustrated in Figure 11.8 , is designed to reflect the different types of potential input in solving balance-beam – type tasks. It has four different groups of input units that receive input about weights and distances for each side of the fulcrum.

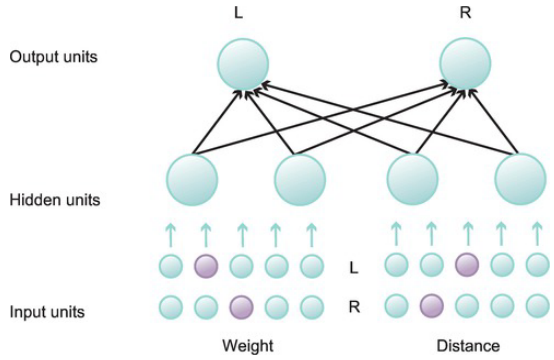


Figure 11.8 The architecture of the McClelland and Jenkins network for the balance-beam problem.

(Adapted from Elman, 1996.)

The information the network gets is actually quite impoverished. One group of input units will get information corresponding to, say, the weights to be found on one side of the beam. Another group of units will get information corresponding to the distances of those weights from the fulcrum. But these are separate pieces of information. The network needs to work out during training that the two groups of units are carrying information about the same side of the balance beam. The network weights are initially set at random.

As we see in Figure 11.8 , the weight units are connected to a pair of hidden units. Likewise for the distance units. There are no connections

between the two pairs of hidden units, but each hidden unit projects to both the output units. The network predicts that the balance beam will come down on the left-hand side when the activation on the left output unit exceeds the activation on the right output unit.

The McClelland – Jenkins network learns by backpropagation. The discrepancy between the correct output and the actual output on given iterations of the task is propagated backward through the network to adjust the weights of the connections to and from the hidden units.

As the training went on, the network went through a sequence of stages very similar to those that Siegler identified in children. The initial training examples showed much more variation in weight than in distance. This was intended to reflect the fact that children are more used to using weight than distance in determining quantities such as overall heaviness. As an artifact of the training schedule, therefore, the network ′ s early discriminations all fell into Siegler ′ s stage 1. As training progressed, the network learned to use distance to solve problems with equal numbers of weights on each side — as per Siegler ′ s stage 2. The final stages of the training saw the network move to Siegler ′ s stage 3, correctly using both weight and distance, provided that the two sides differed only on one dimension but not on both. The McClelland – Jenkins network did not arrive at Siegler ′ s stage 4. But a similar network designed by Jay McClelland did end up showing all four stages.

In sum, progress on the balance-beam problem can be characterized as a steplike progression. Each step seems to involve exploiting a different rule. The most natural way of modeling this kind of learning pattern would be via a model that had these rules explicitly wired into it — exactly the sort of model that would be suggested by the physical symbol system

hypothesis. The qualitative progression between different stages would be explained by the transition from one rule to another.

Neural network models show us, however, that steplike progressions can emerge without explicit rules. There are only two rules explicitly programmed into the network — the activation rule governing the spread of activation forward throughout the network and the backpropagation rules governing the spread of error backward through the network. There is nothing in the network corresponding to the rules in terms of which it might be described. Nor are there any sharp boundaries between the types of learning at different stages, even though its actual performance on the task has a clearly identifiable steplike structure .

11.3 Conclusion: The Question of Levels

The models we ′ ve been looking at have revealed some of the great strengths of artificial neural networks — particularly when it comes to modeling complicated learning trajectories. We have seen how representations in neural networks are distributed across different hidden units and how hard it can be to find any sort of straightforward mapping between what is going on inside the network and the task that the network is performing. In this final section, we will step back from the details of individual neural network models to look briefly at a very important concern that some cognitive scientists have raised about the whole enterprise of neural network modeling.

David Marr ′ s tri-level hypothesis, which we looked at in Section 2.3 , distinguished three different levels at which cognitive scientists study a

given cognitive system. A quick reminder:

- *The computational level* provides a general characterization of the information-processing task that the system is trying to perform.
- *The algorithmic level* identifies a particular algorithm or set of algorithms that can carry out the task identified at the computational level.
- *The implementational level* explains how the algorithm is actually realized in the system.

Think about a Turing machine, for example. Analysis at the computational level identifies the general information-processing task that the machine is performing — for example, computing the arithmetical function of addition. An analysis at the algorithmic level will come up with a specific machine table that will compute this function. And then, when we turn to the implementational level, what we are interested in is how to build a physical system that will run that algorithm.

> **Exercise 11.3** Pick an example of a cognitive phenomenon from anywhere in this book and use it to illustrate in your own words the difference between Marr ′ s three levels of analysis.

The difference between the algorithmic and implementational levels is very important. The implementational level is the level of engineering and machinery, whereas the algorithmic level is the level of discrete information-processing steps, each governed by specific rules. If our Turing machine were a digital computer, the algorithmic-level analysis would

correspond to the program that the computer is running, whereas the implementational analysis would explain how that program is realized in the hardware of the computer.

Physical symbol theorists are typically explicit about the level at which their accounts are pitched. As one would expect, given the emphasis on algorithms and rules for manipulating symbol structures, the physical symbol system hypothesis is aimed squarely as an algorithmic-level account. It is not an engineering-level account of information-processing machinery. Rather, it needs to be supplemented by such an account.

So, how should we think about artificial neural networks? If we have a connectionist model of, say, past-tense learning, is it an algorithmic-level account? Or an account at the implementational level? Do artificial neural networks tell us about the abstract nature of the information-processing algorithms that can solve particular types of cognitive tasks? Or do they simply give us insight into the machinery that might run those information-processing algorithms?

This is important because artificial neural networks will only count as alternatives to physical symbol systems if they turn out to be algorithmic-level accounts. The whole contrast that we have been exploring in the last two chapters between neural network models of information processing and physical symbol system models depends on understanding neural networks at the algorithmic level.

A number of physical symbol theorists (most prominently Jerry Fodor and Zenon Pylyshyn) have used this point to make a powerful objection to the whole enterprise of artificial neural network modeling. They begin by observing that if we think about artificial neural networks at the implementational level, then they are not really an alternative to the

physical symbol system hypothesis at all. They are simply offering models of how physical symbol systems can be implemented.

Moreover, Fodor and Pylyshyn argue, artificial neural networks cannot provide algorithmic-level models. As language-of-thought theorists (see Chapter 4), they think that cognition can only be understood in terms of the rule-governed transformation of abstract symbol structures — a manipulation that is sensitive only to the formal, syntactic features of those symbol structures. And this only works when we have symbol structures composed of separable and recombinable components.

But artificial neural networks simply do not have separable and recombinable components. They have a very different kind of structure. The state of a network at any given moment is fixed by the state of all the units it contains. And because each distinct unit has a range of possible activation levels, there are as many different possible dimensions of variation for the network as a whole as there are units. So, if there are *n* such units, then we can think of the state of the network at any given moment as a point in an *n*-dimensional space — standardly called the *activation space* of the system.

Computation in an artificial neural network is really a movement from one point in the network ′ s activation space to another. But when you think about it like that, it is hard to see how the notion of structure can apply at all. A point on a line does not have any structure. It does not have separable and recombinable components. Nor does a point on the plane (i.e., in two-dimensional space) — or a point in any *n* -dimensional space.

So, Fodor and Pylyshyn ′ s argument is, in essence, a dilemma. Either neural networks contain representations with separable and recombinable components, or they do not. If they do contain such representations, then they are not really alternatives to the physical symbol system hypothesis. In

fact, they will just turn out to be ingenious ways of implementing physical symbol systems. But if, on the other hand, they do not contain such representations, then (according to Fodor and Pylyshyn) they have absolutely no plausibility as algorithmic-level models of information processing. Here is the argument, represented schematically:

1 Either artificial neural networks contain representations with separable and recombinable components, or they do not.

2 If they do contain such representations, then they are simply implementations of physical symbol systems.

3 If they do not contain such representations, then they cannot plausibly be described as algorithmic information processors.

4 Either way, therefore, artificial neural networks are not serious competitors to the physical symbol system hypothesis.

This argument is certainly elegant. You may well feel, though, that it is begging the question. After all, the whole point of the neural network models we have been looking at in this chapter (and Chapter 10) has been to try to show that there can be information processing that does *not* require the type of rule-governed symbol manipulation at the heart of the physical symbol system hypothesis. In a sense, the models themselves are the best advertisement for artificial neural networks as genuine alternative models of information processing — rather than simply implementations of physical symbol systems.

Exercise 11.4 Assess in your own words step 3 in Fodor and Pylyshyn ′ s argument.

In any case, there is no law that says that there is only one type of information processing. Perhaps the physical symbol system approach and the neural networks approach can coexist. It may turn out that they are each suitable for different information-processing tasks. When we explored the language-of-thought hypothesis, for example, we placed considerable emphasis on the role of propositional attitudes such as belief and desire in causing behavior. The interplay of syntax and semantics in the language of thought was intended to capture the idea that beliefs and desires could bring about behavior by virtue of how they represent the world. But the types of tasks we have been looking at in these last two chapters seem very different — closer to perception and pattern recognition than to abstract symbol manipulation. It may turn out that different types of cognitive tasks require fundamentally different types of information processing .

Summary

This chapter first reviewed experiments indicating infants ′ expectations of physical objects and their behavior, which have seemed to many to suggest that infants possess a theory-like folk physics. We then looked at neural network models of infant understanding of object permanence. Without having any rules encoded in it, the network can accurately model infant expectations about when occluded objects are going to reappear from

behind a barrier. It learns to do this because of feedback connections from the hidden units to themselves, which function as a type of memory within the network. After reviewing a neural network model of the balance-beam problem, chapter ended by considering a famous dilemma that Fodor and Pylyshyn have posed for neural network models.

Checklists

Object Permanence in Infancy

1. Traditionally, the perceptual universe of the infant is viewed as a " blooming, buzzing, confusion, " with infants only understanding object permanence (i.e., that objects continue to exist when they are not directly perceived) at around eight months.

2. Recent studies using the dishabituation paradigm have led many developmental psychologists to revise this view and to claim that even very young infants inhabit a highly structured and orderly perceptual universe.

3. Researchers such as Elizabeth Spelke have argued that young infants are able to parse the visual array into objects that behave according to certain basic physical principles.

Neural Network Models of Object Permanence

1. Many cognitive scientists think that object permanence requires explicitly represented physical principles (a primitive folk physics).

2. Munakata ′s neural network model suggests that object permanence might be a matter of having representations of objects that persist when the object is occluded rather than explicitly representing physical principles.

3. This network is a recurrent network, where a feedback loop from the hidden units back to themselves functions as a type of memory.

Modeling the Balance-Beam Problem

1. Studies have explored how young children reason about what will happen to objects of different weights placed on opposite sides of a balance beam at different distances from the fulcrum.

2. Children typically go through a fairly standard learning trajectory, where weight and distance are differently weighted at different stages, before they understand that downward force is a function of both weight and distance.

3. This trajectory can be modeled by a neural network whose hidden units develop associative connections between inputs corresponding to the weights and distances.

4. This suggests that we can think about children ′s developing physical understanding without assuming that they are developing an increasingly sophisticated theoretical understanding of physical principles connecting force, weight, and distance.

The Fodor – Pylyshyn Objection to Neural Network Modeling

1. Fodor and Pylyshyn start with a dilemma: either artificial neural networks contain representations with separable and recombinable components, or they do not.

2. If neural networks do contain such representations, then (they argue) the networks are simply implementations of physical symbol systems.

3. But if they do not contain such representations, then (according to Fodor and Pylyshyn) they cannot plausibly be described as algorithmic information processors.

4. Either way, Fodor and Pylyshyn argue, artificial neural networks are not serious competitors to the physical symbol system hypothesis.

5. *But* — this seems to be begging the question because the central claim of the neural networks is that information processing need not require the type of rule-governed symbol manipulation at the heart of the physical symbol system hypothesis.

Further Reading

The drawbridge experiments described in Section 9.3 were first presented in Baillargeon 1986 and 1987. They have been extensively discussed and developed since then. For a more recent model, see Wang and Baillargeon 2008. Spelke ' s experiments using the dishabituation paradigm are reviewed in many places (e.g., Spelke et al. 1995). A general discussion of habituation methodology can be found in Oakes 2010. Spelke and Kinzler 2007 reviews evidence for infant " core knowledge " in understanding

objects, actions, number, and space. Susan Carey and Renée Baillargeon have extended Spelke ' s " core knowledge " in a number of ways. Summaries can be found in Carey and Spelke 1996, Carey 2009, Baillargeon et al. 2010, and Baillargeon and Carey 2012. Woodward and Needham 2009 is a collection of review articles on infant cognition. Hespos and van Marle 2012 provides a summary pertaining specifically to infants ' knowledge of objects. Cacchione 2013 continues experimental work on infant perception of cohesion. The theory of the " child as little scientist " is engagingly presented in Gopnik and Meltzoff 1997.

One of the first papers exploring connectionist approaches to object permanence was Mareschal et al. 1995. Further, see Mareschal and Johnson 2002. The papers discussed in the text are Munakata et al. 1997, Munakata 2001, and Munakata and McClelland 2003. For a book-length treatment of the power of connectionist approaches in thinking about cognitive development, see Elman et al. 1996 — which also contains a detailed account of the balance-beam network discussed in Section 9.4 (originally presented in McClelland and Jenkins 1991). Plunkett and Elman 1997 is an accompanying workbook with software. Marcus 2003 attempts to integrate connectionist and symbolic approaches. Elman 2005 is another good review. A critical view can be found in Quinlan et al. 2007.

The Fodor and Pylyshyn argument discussed in Section 9.5 can be found in Fodor and Pylyshyn 1988. It has been widely discussed. A number of important papers are collected in Macdonald and Macdonald 1995. See Chapter 9 of Bermúdez 2005 for a general discussion and further references, and see Calvo and Symons 2014 for a book-length overview. Jansen and Watter 2012 proposes a network that the authors claim displays strong systematicity.

Chapter 12

Machine Learning: From Expert Systems to Deep Learning



Overview

12.1 Expert Systems and Machine Learning

Expert Systems and Decision Trees

ID3: An Algorithm for Machine Learning

12.2 Representation Learning and Deep Learning

Deep Learning and the Visual Cortex

12.3 The Machinery of Deep Learning

Autoencoders

Convolutional Neural Networks

12.4 Deep Reinforcement Learning

Overview

This chapter is dedicated to machine learning, one of the hottest topics in contemporary artificial intelligence (AI) and the key to the success of multibillion-dollar corporations such as Google, Facebook, and Amazon.

Section 12.1 introduces expert systems — computer programs that are designed to replicate (and improve on) the performance of human experts in specialized domains, such as identifying diseases or processing credit-card applications. There are different ways of constructing expert systems. One way is to start with human experts and write a program that codifies their collective knowledge. Alternatively, machine learning algorithms analyze large databases of examples to derive rules that can then be used to classify new examples. We illustrate this through ID3, which is an example of a traditional machine learning algorithm.

Traditional algorithms such as ID3 are highly dependent on how their databases are labeled and constructed. They typically require lengthy and complex processes of *feature engineering* to structure the databases. In the subfield of machine learning known as *representation learning*, computer scientists write programs that will do their own feature engineering. And within representation learning as a whole, the greatest advances have come from what is called *deep learning*.

We explore deep learning in Sections 12.2, 12.3, and 12.4. Section 12.2 shows how deep-learning programs typically involve multiple layers of artificial neural networks. These networks are hierarchically organized to extract increasingly complex information from the raw data. The mammalian visual system is an explicit inspiration for this type of construction. Section 12.3 introduces two representative types of neural networks used for deep learning — autoencoders and convolutional neural networks. Finally, Section 12.4 looks at the reinforcement-learning methods behind two of the most spectacular examples of deep learning — the AlphaGo and AlphaGo Zero programs created by Google's Deep Mind research team.

12.1 Expert Systems and Machine Learning

In the field of AI known as *expert systems* research, researchers write computer programs to reproduce (and ideally improve on) the performance of human beings who are experts in a particular domain.

Expert systems programs are typically applied in narrowly defined domains to solve very determinate problems, such as diagnosing specific medical disorders. A well-known expert systems program called MYCIN was developed at Stanford University in the early 1970s. MYCIN was designed to simulate a human expert in diagnosing infectious diseases. It took in information from doctors on a particular patient ' s symptoms, medical history, and blood tests, asking for any required information that it did not already have. It then analyzed this information using a knowledge base of about 600 heuristic rules about infectious diseases derived from clinical experts and textbooks.

MYCIN produced a number of different diagnoses and recommendations for antibiotic treatments. It was able to calculate its degree of confidence in each diagnosis and so present its findings as a prioritized list. Although MYCIN was never actually used as the sole tool for diagnosing patients, a widely reported study at Stanford University ' s medical school found that it produced an acceptable diagnosis in 69 percent of cases. You may think that 69 percent is not very high, but it turns out to be significantly higher than the percentage of acceptable diagnoses from infectious disease experts who were using the same rules and information.

Expert Systems and Decision Trees

Expert systems have become very deeply entrenched in the financial services industry, particularly for mortgage loan applications and tax advice. Most banks these days have online " wizards " that will take mortgage applicants through a series of simple questions designed to lead to a decision on the applicant ' s " mortgage-worthiness. " Mortgage wizards can be represented through *decision trees*. In the simplest form of decision tree, each node corresponds to a question. Each node has several branches leading from it, corresponding to different answers to the question.

Figure 12.1 illustrates a very simple schematic expert system for a loan decision tree. It has two important features. First, it offers a fixed decision procedure. Whatever answers the loan applicant gives to the fixed questions, the decision tree will eventually come up with a recommendation. Second, the presentation in tree form is completely inessential. We can easily convey what is going on in terms of explicit rules, such as the following:

IF income less than $40K, THEN no loan.

IF income greater than $75K AND no criminal record, THEN loan.

IF income between $40K and $75K AND applicant working for one to five years AND credit not good, THEN no loan.

(I have used uppercase letters to bring out the logical structure of the rules.) When the decision tree is written as a computer program, it may well be written using explicit rules such as these.

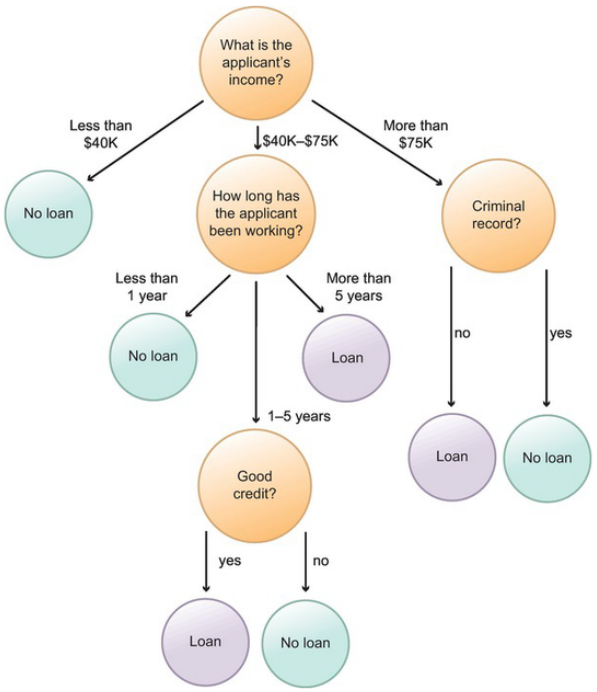


Figure 12.1 A decision tree illustrating a mortgage expert system.

(From Friedenberg and Silverman 2006 .)

The decision tree works because of the questions that are asked at each node. Each question partitions the possibility space so that each branch of the tree leads to a unique outcome (what computer scientists call a *terminal leaf or node*). But where do the questions come from? How does the decision tree get designed?

One possibility would be to ask a team of mortgage loan officers to sit down and work out a decision tree that captures the practices at their bank. This could then be used as the basis for writing a program in a suitable programming language. This would be fine, and it is no doubt how many expert systems programs are actually written (particularly in the mortgage area). But an AI perspective is not very interesting because the real expert system would be the team of mortgage loan professionals. Much more interesting would be a program that was capable of producing its own

decision tree — a program capable of imposing its own structure on the problem and working out what would count as a solution. How would this work?

Suppose that we have a huge database of all the loan decisions that the bank has taken over a long period of time, together with all the relevant information about the applicants — their income, work history, credit rating, and so on. If we can find a way of representing the bank ' s past decisions in the form of a decision tree so that each branch of the tree ends either in the loan being given or the loan being declined, then we can use that decision tree to " process " new applications.

This is a classic example of the type of problem tackled in *machine learning* (a subfield in expert systems research). The challenge is to produce an algorithm that will organize a complex database in terms of some attribute we are particularly interested in (such as an applicant ' s loan-worthiness, in the example we are considering). This process yields a decision tree, which will determine whether or not the attribute holds in a given case (i.e., whether or not the applicant is loan-worthy).

ID3: An Algorithm for Machine Learning

This section explores an influential machine learning algorithm developed by computer scientist Ross Quinlan. Quinlan developed the ID3 learning algorithm while working at the University of Sydney in Australia. He now runs a company called RuleQuest Research, which is commercially marketing updated and more efficient versions of the ID3 algorithm.

Machine learning algorithms such as ID3 only work on databases that take a very specific form. The basic objects in the database are standardly

called *examples* . In the loan application decision tree that we looked at earlier, the examples are loan applicants. These loan applicants can be classified in terms of a certain number of *attributes* . Each example has a value for each attribute. So, for example, if the attribute is *Credit History?* , then the possible values are *Good* and *Bad* , and each mortgage applicant is assigned exactly one of these values. The attribute we are interested in is the *target attribute* . In our example, the target attribute is *Loan* , and the two possible values are *Yes* and *No* . Again, every applicant is either offered a loan or is turned down.

The attributes work to divide the examples into two or more classes. So, for example, the attribute at the top of the decision tree is *Income?* . This attribute divides the loan applicants into three groups. As we move down each branch of the tree, each node is an attribute that divides the branch into two or more further branches. Each branch ends when it arrives at a value for the target attribute (i.e., when the decision is made on whether to give the loan or not).

The ID3 algorithm works by assigning attributes to nodes. It identifies, for each node in the decision tree, which attribute would be most informative at that point, using a statistical measure of informativeness, standardly called *information gain* . Information gain measures how well a particular attribute classifies a set of examples. At each node, the algorithm chooses the remaining attribute with the highest information gain.

The concept of information gain is itself defined in terms of a more fundamental measure called *entropy* . (Warning: You may have come across the concept of entropy in physics, where it features, for example, in the second law of thermodynamics. Entropy is defined somewhat differently in

information theory than in physics, although it is the same basic concept.) Entropy is a measure of uncertainty.

Once we have a formula for calculating entropy, we can calculate information gain relative to a particular attribute, and then we can use that to construct the decision tree. Basically, for each attribute, the algorithm works out by how much the entropy would be reduced if the set were classified according to that attribute. This gives a measure of the information gain for each attribute. Then the algorithm assigns the attribute with the highest information gain to the first node on the tree. The process is continued until each branch of the tree ends in a value for the target attribute.

To see how ID3 works, consider this relatively simple problem — deciding whether or not the weather is suitable for playing tennis. Imagine that, as keen tennis players who seriously consider playing tennis every day, we collect information for 2 weeks. For each day, we log the principal meteorological data and note whether or not we decide to play tennis on that day. We want to use this information to construct a decision tree that we can use in the future.

So, the target attribute is *Play Tennis?* Here are the other attributes with the values they can take:

Outlook?	{sunny, overcast, rain}
Temperature?	{hot, mild, cool}
Humidity?	{high, low, normal}
Wind?	{weak, strong}

And here is our database:

DAY	OUTLOOK?	TEMPERATURE?	HUMIDITY?	WIND?	PLAY TENNIS?
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Even this relatively small database is completely overwhelming. It is very hard to find any correlations between the target attribute and the other attributes. Fortunately, though, this is exactly the sort of problem that ID3 can solve.

First, it finds the attribute with the highest information gain. When ID3 calculates the information gain for all four attributes, the results come out as follows:

Gain (S, *Outlook?*)	=	0.246
Gain (S, *Temperature?*)	=	0.029
Gain (S, *Humidity?*)	=	0.151
Gain (S, *Wind?*)	=	0.048

The information gain is highest for *Outlook?* , so that is the attribute ID3 assigns to the first node in the decision tree. The decision tree looks like the one in Figure 12.2 .

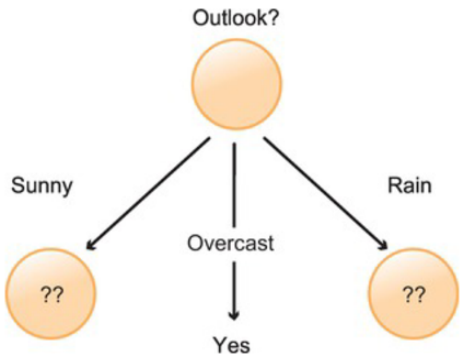


Figure 12.2 The first node on the decision tree for the tennis problem. *Outlook* is the first node on the decision tree because it has the highest information gain.

Each of the three branches coming down from the first node corresponds to one of the three possible values for *Outlook?* . Two of the branches (*Sunny* and *Rain*) lead to further nodes, whereas the middle branch immediately ends.

It turns out that assigning attributes to these two nodes is all that is required for a comprehensive decision tree — that is, for a decision tree that will tell us whether or not to play tennis in any combination of meteorological conditions. The final decision tree is illustrated in Figure 12.3 .

Exercise 12.1 Check that this decision tree works for two of the lines in the original database.

This is a " toy " example. But there are plenty of " real-life " examples of how successful ID3 can be. Here is one.

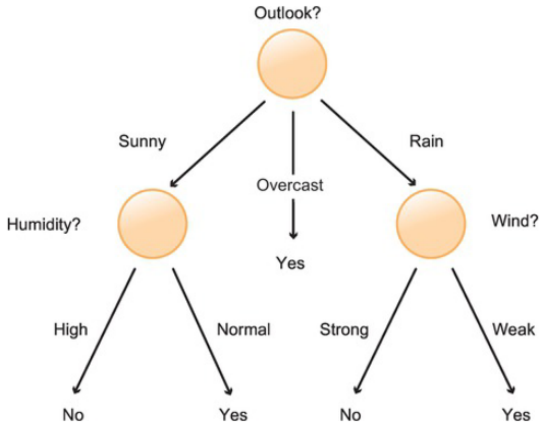


Figure 12.3 The complete decision tree generated by the ID3 algorithm.

In the late 1970s, Ryszard Michalski and Richard Chilausky, two computer scientists at the University of Illinois (deep in the agricultural heartland of America ' s Midwest), used ID3 to devise an expert system for diagnosing diseases in soybeans, one of Illinois ' s most important crops. This is a rather more difficult problem because there are nineteen common diseases threatening soybean crops. Each disease is standardly diagnosed in terms of clusters of thirty-five different symptoms. In this case, therefore, the target attribute has nineteen different possible values, and there are thirty-five different attributes. Many of these attributes also have multiple possible values.

In order to appreciate how complicated this problem is, look at Figure 12.4 . This is part of a questionnaire sent to soybean farmers with diseased crops. It asks for details on a wide range of attributes. Completed questionnaires such as this one were one of the inputs to the initial database. They were supplemented by textbook analyses and lengthy consultations with a local plant pathologist. The total database on which ID3 was trained comprised 307 different examples.

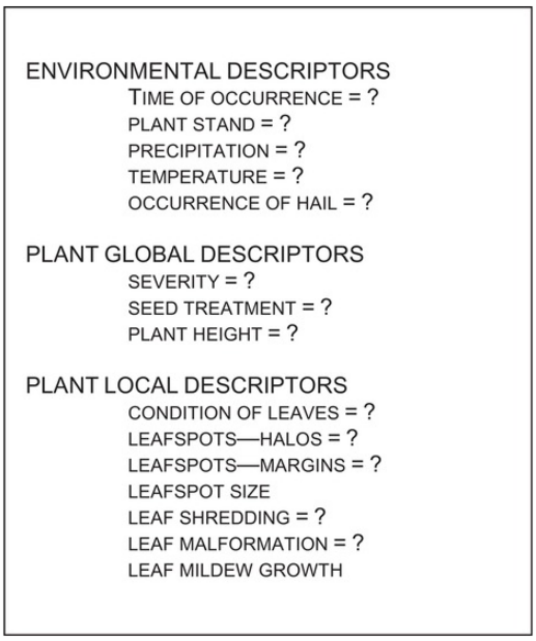


Figure 12.4 A sample completed questionnaire used as input to an ID3-based expert system for diagnosing diseases in soybean crops.

(Adapted from Michalski and Chilausky 1980 .)

Michalski and Chilausky were interested not just in whether ID3 could use the training examples to construct a decision tree. They also wanted to compare the resulting decision tree to the performance of a human expert. And so, they tested the program on 376 new cases and compared its diagnoses to those made by various experts on plant disease (including the author of the textbook that they had originally used to compile the database). As it turned out, the expert system did much better than the human expert on the same 376 cases. In fact, it made only two mistakes, giving it a 99.5 percent success rate, compared to the 87 percent success rate of the human experts.

12.2 Representation Learning and Deep Learning

A decision tree produced by a machine learning algorithm is very different from one reached by pooling and organizing the insights of a team of experts. Imagine the work that it would take, for example, to get a team of tennis coaches to organize the little database from the last section into a simple set of rules that could tell you whether or not to play tennis in different weather conditions. And as we saw from the soybean disease example, expert systems trained with machine learning algorithms can outperform human experts.

But still, think about all the work that has to be done before the machine learning algorithm can do its job. When Michalski and Chilausky used ID3 to train an expert system to diagnose diseases in soybean crops, they started with a database that was already highly organized. As shown in Figure 12.4 , the examples in the database were classified in terms of highly complex attributes, including different types of information about the plant ′ s leaves, fruit pods, and seeds, as well as general facts about the weather.

In a sense, therefore, much of the work had already been done before the machine learning algorithm had even gotten going. And so, you might think, all of this preliminary work really reduces the extent to which the algorithm can properly be described as learning. Applying a machine learning algorithm such as ID3 to a database depends crucially on how the examples in the database are labeled and categorized. And that is an activity that is typically performed by human experts.

In machine learning, this is called *feature engineering* . It is not a trivial task because data sets can typically be labeled in many different ways, and

some ways are much easier to work with than others.

For a simple example, consider elementary arithmetic. You probably learned some simple algorithms for multiplying numbers at a very early age. These algorithms are so familiar that it is easy to forget how dependent they are on representing numbers in base 10. You should have no difficulty multiplying 43 by 17. But now try doing the same thing in Roman notation (where you now have to multiply XLIII by XVII). Or in binary notation (where the task is now to multiply 101011 by 10001). Computers typically work in binary, but you will probably struggle with binary multiplication.

For another example, look at the two diagrams in Figure 12.5 . Imagine that you have a database of examples that you are trying to divide into two categories. One way to do that is to use the information in the database to plot all of the examples on a scatterplot graph and then see if you can write down the equation of a line that separates the two groups. But you have a choice of coordinate systems. In the left-hand diagram, the examples are all plotted using Cartesian coordinates (i.e., in terms of their position on an ordinary x -axis and y -axis). As you can see, the examples do clearly fall into two groups, but because the groups are concentric bands, you will need to write the equation for a circle. Exactly the same data are presented in the right-hand diagram, but here, they are plotted using polar coordinates (e.g., in terms of their distance from a reference point in the middle of the two concentric circles and their angle from the reference direction of straight upward). Now the two categories of examples are separated by a straight line, and the equation is much easier to write.

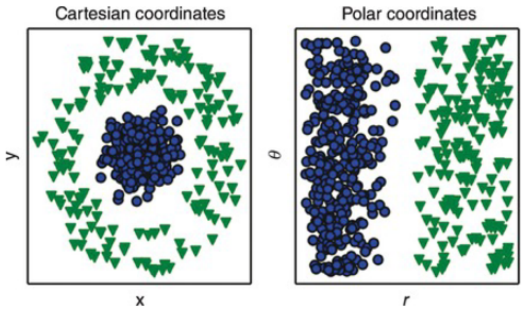


Figure 12.5 Different ways of distinguishing two groups in a database of examples. The left-hand representation uses Cartesian coordinates, whereas the right-hand representation uses polar coordinates. The right-hand representation makes it much easier, for example, to write the equation for a line separating the two groups.

(From Goodfellow et al. 2016 .)

Exercise 12.2 Can you think of another example where the difficulty of solving a problem varies according to how the problem is formulated?

These two illustrations are both examples of feature engineering. For the multiplication example, the key feature of the numbers being multiplied is how they can be represented in our ordinary base 10 notation (what we more often call *decimal notation*) as opposed to how they can be represented in base 2 notation (binary) or the Roman version of base 10 notation. And for the scatterplot example, the key features are how the examples can be represented in terms of their distance and angle from a reference point and reference direction.

Exercise 12.3 Think of another example of feature engineering .

Standard machine learning algorithms such as ID3 can only get to work once the feature engineering has been carried out, typically by programmers and other human experts. And this feature engineering is often the most complicated and time-consuming part of the process. It is not hard to see the advantages of designing programs and algorithms that would be able to do this feature engineering for us. The technical name for this is *representation learning* or *feature learning* .

This brings us to *deep learning* , which will occupy us for the rest of this chapter. Deep learning is one of the hottest topics in contemporary AI and computer science, pioneered by university researchers and by teams at major tech companies (such as the Deep Mind team associated with Google). Deep-learning algorithms have been responsible for important theoretical advances in areas of AI that had traditionally been viewed as the most challenging and intractable. These include computer vision and natural language processing, areas where computer scientists have long struggled to make anything but the most basic advances.

Deep learning has also had practical applications across a wide range of areas. These include software for reading handwriting (not just in languages with alphabets, such as English, but also in Chinese) and hence for processing handwritten documents, such as checks, for example. Deep learning has been extensively applied to speech recognition. As a result, voice-based AI, such as Amazon ' s Alexa, is typically powered by deep-learning algorithms. Deep learning is also a key component in the technology behind self-driving cars.

As you might expect, there are many different understandings of deep learning. But here is a very useful characterization from a review article in the journal *Nature* by three pioneers of deep learning — Yann LeCun, Yoshua Bengio, and Geoffrey Hinton:

> Representation learning is a set of methods that allows a machine to be fed with raw data and to automatically discover the representations needed for detection or classification. Deep-learning methods are representation-learning methods with multiple levels of representation, obtained by composing simple but non-linear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level.
>
> (LeCun et al. 2015 : 436)

LeCun and colleagues define deep learning as a special type of representation learning. We will unpack this general definition, focusing in particular on the general idea of multiple and hierarchical levels of representation. We will then look at specific architectures and algorithms for deep learning.

Deep Learning and the Visual Cortex

Deep-learning systems are typically constructed from multiple layers of artificial neural networks. Although these networks are descendants of the connectionist models explored in earlier chapters, the design of the individual networks and the overall architecture of the systems are very different.

Deep-learning theorists often explicitly appeal to the mammalian visual cortex as a model and inspiration. This is for two reasons. First, the mammalian visual cortex only works because it is capable of representation learning. And second, the general design of the mammalian visual cortex is a blueprint for designing representation-learning algorithms, particularly because it is hierarchically organized, with different areas interpreting progressively more complex and structured representations.

The visual system confronts a classic problem of representation learning. It has to take a complex pattern of unstructured stimuli in the visual field and interpret it into representations that can then serve as input to more complex cognitive functions, such as object recognition. It is a natural representation-learning system. By analogy, an artificial representation-learning system has the parallel task of taking complex and unstructured raw data and transforming the data into representations that can serve as inputs to systems carrying out more complex tasks, such as classification.

The organization and functioning of the mammalian visual cortex are relatively well understood. There is a strong consensus among cognitive neuroscientists that information in the visual cortex is processed hierarchically. Information flows through a progression of different areas, each generating representations of increasing complexity.

The first station is the *lateral geniculate nucleus* (LGN), which receives input directly from the retina. Because neurons in the LGN have similar receptive fields to neurons in the retina, the LGN functions as a relay. It projects to area V1 (the primary visual cortex, which maps more or less onto the anatomically defined area known as the *striate cortex* because of its conspicuous stripe). This is where “ information processing ” proper begins.

The neurons in V1 are sensitive to low-level features of the visual field, such as orientation and direction of movement. Like the LGN, V1 is retinotopically organized — that is, neighboring regions of the visual field are represented by neighboring regions of V1. In broad outline, V1 takes the retinotopic map coming from the LGN and filters it in a way that accentuates the edges and contours. This is the first step in parsing the raw data arriving at the retina into objects.

Area V1 projects to area V2 (also known as the *secondary visual cortex*). The neurons in V2 are tuned to the same features as the neurons in V1, as well as to more complex features, such as shape and depth. V2 is much less well understood than V1, but the neurons there are sensitive to multiedge features, such as complexes of edges with different orientations.

As we saw in Mishkin and Ungerleider ′ s two-visual-systems hypothesis in Section 3.1 , there are two distinct neural pathways for information leaving area V2. The ventral pathway and the dorsal pathway are each believed to be responsible for different types of visual processing:

- The *ventral pathway* , thought to be responsible for object identification and recognition, goes from V2 to V4 and then onward to areas in the inferior temporal cortex.
- The *dorsal pathway* , more focused on representations relevant to action, goes from V2 to areas V5 and V6 (also known as the *middle temporal* [MT] and *dorsomedial* [DM] areas).

We are currently interested primarily in the ventral pathway because we are interested in how the visual system solves the representation-learning problem of mapping unstructured stimuli in the visual field onto representations of objects in the distal environment — representations that

can serve as inputs for more complicated forms of classification and information processing.

Area V4 takes the representations from V2, which contain primitive representations of shapes and some information about depth, and then uses those representations to construct further representations that incorporate more information about figure/ground segmentation, as well as about colors. Area V4 is thought to be the first stage in early visual processing that is modulated by attention. Representations leaving V4 then go to the inferior temporal cortex (ITC), where there are various areas specialized for sensitivity to high-level patterns. Areas in the ITC include the fusiform face area (FFA), thought to be specialized for face recognition, and the fusiform body area (FBA), believed to be dedicated to identifying the human body and body parts.

The general pattern in the visual cortex is increasing determinacy of representation, with different areas in each pathway using the outputs of earlier areas to identify increasingly complex features and complexes of features. The different areas in the ventral pathway form a hierarchy, which is illustrated in Figure 12.6 .

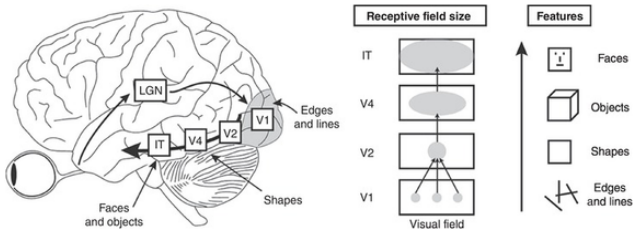


Figure 12.6 An illustration of hierarchical visual processing. Stimuli are processed in a series of visual areas, which process features of increasing complexity.

(From Figure 1 in Manassi et al. 2013 .)

This type of hierarchical information processing is (by design) exactly what we find in many deep learning systems. Deep learning often involves multiple layers of information processing, each building on the outputs of earlier layers.

The general reasoning behind this approach is explained clearly by LeCun et al. (2015 : 439):

> Deep neural networks exploit the property that many natural signals are compositional hierarchies, in which higher-level features are obtained by composing lower-level ones. In images, local combinations of edges form motifs, motifs assemble into parts, and parts form objects. Similar hierarchies exist in speech and text from sounds to phonemes, syllables, words and sentences. The pooling allows representations to vary very little when elements in the previous layer vary in position and appearance.

If, as the authors point out, many natural signals are hierarchically organized, then a deep-learning system has to learn two things. First, it has to learn how to extract the bottom-level features. And then, second, it has to learn how to compound them to yield an accurate representation of the input in much the same way as the mammalian visual system has learned through evolution to parse the light energy hitting the retina into representations of objects.

This, in turn, means solving what LeCun and colleagues call the *selectivity/invariance problem* , which they illustrate with a canine example. Samoyeds are beautiful white dogs from the husky family. Imagine that you are trying to construct a system that will identify Samoyeds and distinguish them from other animals. Because Samoyeds have long fur, it is easy enough

to distinguish them from short-haired dogs such as greyhounds. But what about distinguishing Samoyeds from their distant relative, the wolf?

The system needs to develop a highly *selective* representation of a Samoyed, one that will pick up on the relatively small differences between a Samoyed and a white wolf. But at the same time, the representation needs to be *invariant* in a way that ignores the conspicuous differences between what the animal looks like from different angles and in different postures. To illustrate the problem, think about how similar a Samoyed and a white wolf would look if they were both sitting down and viewed face on, and compare this to how different two Samoyeds would look if one was lying down and viewed from above while the other was running and viewed from in front.

Deep neural networks can solve the selectivity/invariance problem because they are constructed as multilayered stacks of simple networks. Each simple network performs a mapping from input to output, with the output serving as the input to the next network in the stack (just as the output of area V1 in the visual cortex is the input to area V2). Each of those mappings codes the input in terms of increasingly complex features (just as area V4 takes the primitive shapes in the output from area V2 and segments them into figure and ground to enhance the representation of depth).

The details of how this process works are extremely technical, but there are two fundamental types of neural networks that feature in many deep-learning systems and that can be explained in a relatively nontechnical way. The first type is called an *autoencoder* . The second type is called a *convolutional neural network* . We turn to them in the next section .

12.3 The Machinery of Deep Learning

In this section, we ′ ll start by looking at a type of neural network called an *autoencoder* . Autoencoders are a very nice example of how a network can learn to extract features from unstructured raw data. Then we turn to what was probably the most important innovation in the deep-learning revolution, namely, the development of convolutional neural networks (ConvNets). ConvNets are a special kind of feedforward neural network, particularly suited to processing data organized in a grid-like format (a photograph, for example, or the human visual field).

Autoencoders

Think about the party game of charades. When you play charades, your job is to take a word or phrase and then mime it out so that the other players can guess what you started with. In other words, you take a signal as input and then represent it in such a way that the original signal can be recovered as output. Your mime performance is the representation, and the other players ′ guess is the output. You have been successful if the output matches the input (i.e., if the other players come up with the word you started with).

This is pretty much what an autoencoder does. It is a single neural network that learns to match output to input. This might seem a very trivial thing to aim for — surely all the network has to do is repeat the input, which would guarantee success? But the point of an autoencoder is that it has constraints built into it that prevent it from solving the problem so simply — just as the rule in charades that you are only allowed to mime prevents you from simply telling the other players the word or phrase you started with. These constraints effectively force the network to find a simplified

representation of the input. The representation needs to be simplified enough to satisfy the constraints built into the network but not so simplified that the network cannot recover the original signal.

An autoencoder has two components. One component *encodes* the original signal, whereas the other *decodes* it. Both encoding and decoding are done by hidden layers. The aim of the network is to minimize the difference between input and output, and it is trained through some form of error-minimization algorithm, such as backpropagation. An autoencoder will typically have built into it some type of bottleneck that forces the network to find a more compressed way of representing the input.

So, for example, if the network has 100 input units and 100 output units, then the bottleneck might be a hidden layer with only 50 units. This constraint forces the network to compress an input signal that has 100 dimensions of variation into a representation that only allows variation in 50 dimensions. The network has to find key features of the input signal that will allow the encoded representation to be decoded back to something close to the original signal. Looking at it like this explains why feature learning is sometimes called *dimensionality reduction* .

There are two illustrations of autoencoders in Figure 12.7 . The left-hand diagram is a schematic wiring diagram for an autoencoder neural network, whereas the second shows what such a network might actually do.

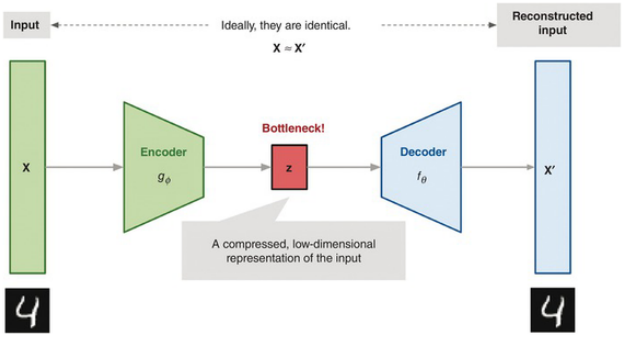


Figure 12.7 Illustration of how an autoencoder compresses and then decompresses a signal.

(Reproduced by permission of Lilian Wengweng.)

Autoencoders can be used to *pretrain* deep neural networks as an alternative both to starting with a random assignment of weights and to using some form of supervised-learning process (where the network is given explicit feedback). Here is how it works. The autoencoder learns to compress its input signal in such a way that the signal can be decoded fairly accurately. This is a form of feature learning. The network is picking up on salient features in the input and using them to encode the original signal. Once the network has learned to do this, it will encode any relevantly similar input signal in the same way. So, now the encoding component of the autoencoder can be built into a deep neural network as one of its layers.

Imagine that we are trying to build a deep neural network that approximates visual processing in the ventral stream (not as fanciful as it sounds — as we ′ ll see shortly, this has actually been done). So, we need to build into it a network that will carry out edge detection. One possibility would be to train an autoencoder on images relevantly similar to those serving as inputs to area V1.

As the autoencoder learns to compress and decompress those images, it effectively learns to encode images as complexes of edges. We can then bypass the decoding component of the autoencoder and wire the bottleneck hidden layer up so that it becomes the input into the layer of the deep neural network corresponding to area V2. And then we might repeat the same process with an autoencoder that would learn to carry out V2-type feature detection, developing sensitivity to multiedge features.

> **Exercise 12.4** Explain in your own words the basic idea behind an autoencoder.

In practice, there are better and more effective ways of constructing deep and multilayer neural networks (e.g., by using the convolutional networks that we will look at next). But autoencoders are a good illustration of some basic deep-learning principles.

Convolutional Neural Networks

For a long time, computer vision was a final frontier for AI. It seemed almost impossible to design an artificial system to perform even a tiny fraction of the complicated visual information processing that a human child or adult performs almost every waking second of their lives. That has all changed in the last few years. The capabilities of the best modern machines are now not far short of their human models. This, in turn, is making possible a tsunami of technological advances, from facial-recognition software in cell phones to self-driving cars. ConvNets are playing a key role in this technological transformation.

The power of ConvNets was first revealed at a machine-vision competition in 2012. The ImageNet Large-Scale Visual Recognition Challenge uses the massive ImageNet database created by Li Fei-Fei and Kai Li, on the faculty at Stanford and Princeton, respectively. ImageNet is actually derived from a lexical database called WordNet, which classifies English words and expressions into classes of synonyms called *synsets* . The goal of WordNet is really to provide an exhaustive catalog of all the concepts expressible in English. There are around 100,000 synsets in WordNet (around 80 percent associated with nouns and noun phrases), and ImageNet contains on average 1,000 images for each synset — and so contains around 10 million images in total.

Computer-vision algorithms entered into the ImageNet competition had two different tasks to perform on a test database of images. The first task was to decide, for an evaluation set of image – category pairs, whether each image actually contained an example of an object from the relevant category (whether the image contained a washing machine, for example). The second task was to identify a specific object within an image and place a box around it (e.g., to identify a screwdriver centered on a specific pixel in the image).

The ImageNet competition started in 2010. Two years later, in 2012, a team from the University of Toronto entered a program called SuperVision. The team was led by Geoffrey Hinton, who had been an early pioneer of backpropagation learning for neural networks and then, in the early 2000s, was one of the researchers at the Canadian Institute for Advanced Research (CIFAR) responsible for the development of deep learning.

SuperVision was wildly successful. Its error rate on the test data set was only 16.4 percent. In comparison, the programs that won the competition in

2010 and 2011 had error rates of 28.2 percent and 25.8 percent, respectively, and the second-place program in 2012 had an error rate of 26.2 percent.

SuperVision ' s secret? It was one of the first computer-vision programs to use a convolutional neural network. The success of ConvNets in object recognition and related tasks has been so compelling that almost all modern computer-vision programs use them. And in fact, it did not take long to use ConvNets to develop programs much more successful even than SuperVision. The winner of the 2014 ImageNet competition, a program created by engineers from Google called GoogLeNet, had an error rate of only 6.7 percent.

ConvNets use a different type of mathematics from traditional neural networks. In traditional neural networks, activation spreads through the network via a form of matrix multiplication. This works because you can represent the activation level of neurons in each layer and the network weights as arrays of numbers (matrices). Activation spreads through the network as a function of the activation levels of individual units and the weights that those individual units have. Matrix multiplication is the mathematical tool for combining arrays of numbers.

Convolution is a more complex mathematical operation. Basically, it creates a type of filter that functions as a localized feature detector. Imagine a grid composed of 10,000 pixels (organized in a square with dimensions 100 × 100). A convolutional layer will successively examine subgrids of this grid to filter out everything except the particular feature that it has been trained to detect. You can think of the process as a little square window (of size, say, 5 × 5) scanning the grid to detect edges, for example.

A ConvNet is a network that has at least one convolutional layer. According to *Deep Learning* , an influential textbook written by Ian

Goodfellow, Yoshua Bengio, and Aaron Courville, ConvNets are designed with three characteristic features that together make them highly suited to processing data that has a grid-like organization (such as an image, for example). Here they are (in slightly different terminology):

- Sparse connectivity
- Shared weights
- Invariance under translation

We will look at these in order.

Sparse Connectivity

Like many traditional neural networks, ConvNets are feedforward. So, activation flows through the network, without any feedback loops. And they have multiple hidden layers. But the wiring is crucially different. In a traditional neural network, each unit in a given layer is connected to all the units in the previous layer and to all the units in the following layer. In other words, each unit is *fully connected* .

In contrast, the hidden layers in ConvNets are typically not fully connected. Each unit is wired to receive inputs only from a small number of units in the preceding layer, and it projects forward only to a small number of units in the next layer. This means that units are, in effect, specialized to respond to local regions of the input to any given layer. So, thinking about this general idea as it might apply to machine vision, you can think of individual units as specialized to respond to particular regions of the visual array.

The sparse-connectivity feature of ConvNets is directly inspired by the mammalian visual system. As we saw earlier, information processing in the early visual system often takes as input retinotopic maps that are more or less isomorphic to the patterns of light energy reaching the retina. Individual neurons in, say, area V1 have a receptive field that only covers a small part of that retinotopic map.

Shared Weights

In ConvNets, as in traditional neural networks, every unit has a weight, and learning takes place through changes in the weights. A unit ′s weight is really a measure of how much it contributes to the overall behavior of the network — a low weight indicates a small contribution, and increasing the weight increases the significance of the unit to the network.

Characteristically, all the weights in a given layer of a traditional network are completely independent of each other. Each unit learns on its own, as it were. The backpropagation learning algorithm, for example, works by assigning to each individual unit a share in the responsibility for the overall degree of error. It then adjusts the weights of each unit to reduce the error attributable to that unit, which has the result of reducing the error in the network as a whole.

In ConvNets, on the other hand, there are dependencies between the weights of different units. A ConvNet might have multiple units with the same weight. Or it might have units whose weights are a function of the weights of other units. This feature adds further to the efficiency of ConvNets relative to traditional networks. Sparse connectivity means that ConvNets have fewer connections. Fewer connections result in fewer

computations as activation flows through the network. Reducing the number of different weights involved reduces the overall number of computations still further.

Invariance under Translation

The last distinguishing feature of ConvNets arises from how the shared weights are set up. To understand it, think again about hierarchical information processing in the visual system. The various areas in the ventral stream generate a succession of maplike representations of the distal environment, with each map building on earlier maps and adding more complex features. This process works because the structure of the original map is preserved through a series of successive transformations.

The visual system often has to process retinal images derived from similar scenes — or from different perspectives on the same scene. It would be desirable for similarities in structure to be preserved across different episodes of vision. For an example, consider how the object-recognition systems in the inferior temporal cortex might recognize that the object you are now looking at is the same object you were looking at from a different angle a couple of minutes ago. The object-recognition system needs to pick up on both similarities and differences. It needs to be sensitive to similarities to register that this is indeed the same object, but at the same time, it needs to be sensitive to differences in the object ' s visual appearance due to the changed angle.

For the object-recognition system to be sensitive to similarities and differences in this way, it must be the case that similar sorts of changes in objects produce similar sorts of changes in the images that represent them.

So, for example, whenever an object moves a certain distance to the left, the image of the object will move a certain distance to the left in the map that the visual system is building of the distal environment. This is the property of invariance under translation. ConvNets have the property of invariance under translation because changes in the input are systematically reflected in the output.

Exercise 12.5 Write down brief definitions in your own words of the three characteristic features of ConvNets.

To conclude, then, ConvNets incorporate a range of features that make them particularly well suited to processing information that comes in a grid-like format. That is why they are so successful at object recognition and why they now dominate machine vision. But many types of information can be presented in a grid-like format. Linguistic information is an example. A sentence, after all, is really just a one-dimensional grid of words. And so, it is not surprising that ConvNets have also been successfully applied in natural language processing. They have proved surprisingly good at speech recognition and semantic parsing, even with no linguistic information coded into them .

12.4 Deep-Reinforcement Learning

One of the main themes emerging from our discussion of deep learning is how biologically inspired it is. On a large scale, deep neural networks are typically constructed in multiple layers to implement hierarchical

information processing explicitly modeled on information processing in the primate visual system. On a smaller scale, the sparse connectivity and shared weights that we find in convolutional neural networks are much more neurally plausible than the full connectivity and atomistic learning that we find in traditional neural networks. We turn now to another example of biologically plausible deep learning.

As was extensively covered in the national and international press, the AlphaGo program created by Google ′ s Deep Mind research group has been spectacularly successful at beating the world ′ s leading experts at the game of Go. At the Future of Go summit in 2017, AlphaGo defeated Ke Jie, the number-one player at the time, building on previous victories in 2016 (watched by tens of millions of people worldwide) over Lee Sedol (see Figure 12.8), also one of a handful of top human Go players. These victories were widely recognized as historic achievements for AI. Go is even more complex than chess because the games are longer and there are many more legally permissible moves available (250 for Go, as opposed to 35 for chess).



Figure 12.8 A move in the Google DeepMind challenge between AlphaGo and Lee Sedol in 2016.

(Handout/Getty Images.)

AlphaGo was initially trained on a massive database of 30 million moves from an online server. The training was a form of supervised learning, with AlphaGo receiving explicit feedback on how successful it was. Once AlphaGo had achieved a relatively high level of playing strength, the training shifted to reinforcement learning.

Reinforcement learning is not like any of the types of machine learning we have looked at up to now. Reinforcement learning differs from supervised learning. The network does not receive explicit feedback on how well it is doing, and it is not learning from a labeled training set. But nor is reinforcement learning a form of unsupervised learning, which is what we find in convolutional neural networks. In unsupervised learning, networks learn to detect patterns in data. An unsupervised network might learn to become an edge detector, for example.

Reinforcement learning is distinctive because (unlike unsupervised learning) it *does* depend on a feedback signal. But at the same time (unlike supervised learning), the feedback signal does not tell the network exactly what it has done wrong. Instead, reinforcement learning is driven by a reward signal. The job of a network incorporating a reinforcement-learning algorithm is to maximize the reward. But it is not told how to do that. It has to work out for itself which outputs are most profitable so that it can repeat and/or adapt them. In a slogan, *supervised learning proceeds through instruction, whereas reinforcement learning proceeds through reward* .

Because the feedback in reinforcement learning is indirect, the network has to engage in a degree of trial and error. It will not increase its reward simply by repeating what has worked in the past. It needs to try new strategies to see whether they work better than current ones. So, successful

reinforcement learning involves striking a balance between exploring new strategies and sticking to tried-and-tested strategies.

Reinforcement learning in AlphaGo was achieved by getting the network to play games of Go against former versions of itself (selected at random from earlier iterations of the supervised-learning process). The reward signal was very simple: +1 for a win and −1 for a loss. AlphaGo then used a learning algorithm called *stochastic gradient descent* . Stochastic gradient descent works to minimize a cost function, where the cost is the negative reward that comes from losing. The algorithm changes the weights in the network so that the negative reward eventually becomes a positive reward.

So, AlphaGo ' s training was a mixture of supervised learning and reinforcement learning. Human participation was crucial to the supervised-learning phase and then dropped out of the picture for the reinforcement-learning phase. You might be wondering, then, whether it is possible to cut the human piece out of the equation completely. Could a network be trained from scratch to play Go, using only reinforcement learning and without any access to databases of position and games or any kind of expert knowledge?

Well, a few months after officially retiring AlphaGo, the team at DeepMind announced a new version of the program, which they called AlphaGo Zero (so-called because it incorporates zero supervised learning). The only thing it knows about Go when it starts is the rules, and so it starts off playing completely at random. Using an updated reinforcement-learning algorithm, the network modifies its weights by playing against itself.

AlphaGo Zero learned very quickly to beat previous iterations of AlphaGo. After three days (and 4.9 million games played against itself), it was able to beat the version of AlphaGo that had defeated Lee Sedol in 2016

— and it was a decisive victory, by 100 games to 0. Within forty days, it was playing Go well enough to beat all existing versions of AlphaGo. Moreover, it was computationally much more efficient than AlphaGo, consuming only a fraction of the power and using only four of Google ' s proprietary tensor processing unit (TPU) chips.

Reinforcement learning is significant from a broader AI perspective. AI is not just about building powerful programs and algorithms that will harness the power of supercomputers to solve difficult problems. AI is also about developing techniques that are recognizably intelligent. And an important index of intelligence in computer programs is that they solve problems in ways similar to how we humans solve them (as opposed to beating them into submission by sheer computational power).

It is important, therefore, that reinforcement learning is the most widespread and powerful form of learning in the animal kingdom. Behaviorists may not be correct that all learning is ultimately reinforcement learning (or, in different terminology, *instrumental conditioning*). But there is no doubt that many, if not most, intelligent behaviors in humans and animals are the result of an exquisite sensitivity to action – reward contingencies. By incorporating reinforcement learning, therefore, deep neural networks can make a strong claim to being intelligent problem solvers.

In fact, the biological plausibility of reinforcement-learning algorithms goes even deeper. There is intriguing evidence from neuroscience that important learning mechanisms in the brain incorporate elements that we also find in reinforcement-learning algorithms. The key findings here come from studies of the neurotransmitter *dopamine* (a neurotransmitter is a chemical released by neurons to send signals to other neurons). Dopamine is

known to be deeply implicated in reward processing (and for that reason has been much studied in the context of addiction).

Pioneering studies of dopaminergic neurons during the 1990s led to the *reward-prediction error hypothesis* of dopamine activity. According to this hypothesis, first proposed by Read Montague, Peter Dayan, and Terrence Sejnowski in 1996, the job of dopaminergic neurons is to fire when the actual reward differs from the expected reward. In other words, to tie this back to machine learning, dopaminergic neurons provide a reward-based error signal, which is exactly what is needed for reinforcement learning. Moreover, although the details are too technical to go into here, it turns out that there are striking parallels between the firing behavior of dopaminergic neurons and how the reward signals are processed in widely used gradient-descent learning algorithms .

Summary

Section 12.1 showed how machine learning algorithms can construct expert systems that can in turn reproduce (and in some cases improve on) the performance of human experts in tasks such as medical diagnosis and processing of mortgage applications. We looked at a specific example in the ID3 algorithm. In Section 12.2 , we looked at the limits of traditional machine learning and, in particular, its dependence on lengthy and complex processes of feature engineering to label and organize databases. Representation learning is the subfield of machine learning dedicated to constructing algorithms that will carry out feature engineering on raw data. The most successful examples of representation learning have come from

deep-learning neural networks, which we looked at in Sections 12.2 , 12.3 , and 12.4 .

Deep-learning networks are hierarchically organized on the model of the mammalian visual cortex. A deep-learning network is made up of layers of individual neural networks, each responsible for extracting more complex features from the original raw data. We looked at two examples of neural networks that can feature in the layers of deep-learning networks — autoencoders and convolutional neural networks. In Section 12.4 , we looked at a further type of learning — reinforcement learning, which is neither supervised learning nor unsupervised. We saw how reinforcement-learning algorithms made possible two of the most spectacular examples of deep learning — the AlphaGo and AlphaGo Zero programs created by Google ′ s Deep Mind research team.

Checklists

Expert Systems and Machine Learning

1. Expert systems are designed to reproduce the performance of human experts in particular domains (e.g., medical diagnosis or financial services).

2. They can typically be represented as decision trees.

3. One problem studied in the field of machine learning is developing an algorithm for generating a decision tree from a complex database.

The ID3 Machine Learning Algorithm

1. ID3 looks for regularities in a database of information that allow it to construct a decision tree.

2. The basic objects in the database are called *examples* . These examples can be classified in terms of their *attributes* . Each attribute divides the examples up into two or more classes.

3. ID3 constructs a decision tree by assigning attributes to nodes. It assigns to each node the attribute that is most informative at that point.

4. Informativeness is calculated in terms of *information gain* , which is itself calculated in terms of *entropy* .

Representation Learning and Deep Learning

1. Problem solving is often highly dependent on how data sets are labeled through *feature engineering* , typically carried out by human experts.

2. Traditional machine learning algorithms, such as ID3, work on databases that are already highly organized and labeled.

3. Representation learning is the subfield of machine learning dedicated to designing algorithms that will do their own feature engineering on raw data.

Deep learning is a special kind of representation learning, carried out by multilayered artificial neural networks, hierarchically

organized to extract increasingly complex features from input raw data.

1. The hierarchical organization of deep-learning networks is inspired by the hierarchical organization of the mammalian visual system.

2. An autoencoder is an example of a deep-learning network that can learn to extract features from raw data.

3. Autoencoders compress raw data through a bottleneck so that the data can subsequently be decoded.

4. Convolutional neural networks (ConvNets) are particularly well suited to feature engineering in data presented in a grid-like format.

5. ConvNets have been spectacularly successful in machine vision.

6. ConvNets feature (a) sparse connectivity, (b) shared weights, and (c) invariance under translation.

Reinforcement learning is a type of machine learning distinct both from supervised learning (in which networks are given explicit feedback on what they have done wrong) and unsupervised learning (such as representation learning).

1. In reinforcement learning, networks receive a reward signal rather than an explicit error signal.

2. Reinforcement-learning algorithms aim to maximize the expected reward, so they need to engage in trial and error to discover new reward-generating strategies.

3. Reinforcement learning is key to the two Go-playing networks, AlphaGo and AlphaGo Zero, developed by Google ′ s Deep Mind

research team.

4. Reinforcement learning is biologically plausible, both because reinforcement learning is the most common type of learning in the animal kingdom and because of intriguing parallels between the firing activity of dopaminergic neurons and how reward signals are processed in widely used learning algorithms.

Further Reading

Much of the literature in traditional machine learning is very technical, but there are some accessible introductions. Haugeland 1985 and Franklin 1995 remain excellent introductions to the early years of AI research. Russell and Norvig 2009 is much more up to date. Also see Poole and Mackworth 2010 , Warwick 2012 , and Proudfoot and Copeland ' s chapter on artificial intelligence in *The Oxford Handbook of Philosophy of Cognitive Science* (Margolis et al. 2012). Medsker and Schulte 2003 is a brief introduction to expert systems, whereas Jackson 1998 is one of the standard textbooks. The *Encyclopedia of Cognitive Science* also has an entry on expert systems (Nadel 2005). See the online resources for a very useful collection of machine learning resources.

The application of ID3 to soybean diseases described in Section 12.2 was originally reported in Michalski and Chilausky 1980 . The database for the tennis example explored in Section 12.1 comes from Chapter 3 of Mitchell 1997 . Wu et al. 2008 describes more recent extensions of ID3, including C4.5 and C5.0, as well as other data-mining methods.

LeCun et al. 2015 is an article-length introduction to deep learning and the source of the definition of deep learning in Section 12.2 . For a more detailed discussion of the theoretical background, see the textbook *Deep Learning* (Goodfellow et al. 2016), which is also available online for free by agreement with the publisher at www.deeplearningbook.org . For a direct deep-learning model of the visual cortex, see Cadieu et al. 2014 . For the SuperVision program entered in the ImageNet competition in 2012, see Krizhevsky et al. 2012 . Lake et al. 2017 argues that deep-learning approaches cannot model intelligence unless they are supplemented by causal models.

The Deep Mind research group at Google has a website at www.deepmind.com and has published articles in *Nature* on AlphaGo (Silver et al. 2016) and AlphaGo Zero (Silver et al. 2017). For a more general discussion of reinforcement learning, the classic textbook is Sutton and Barto 1998 , with a second edition published in 2018 and available from Richard Sutton ' s website at www.incompleteideas.net/book/the-book-2nd.html . Chapter 15 of the second edition covers the neuroscience dimension of reinforcement learning. Pioneering papers on the reward-prediction error hypothesis include Montague et al. 1996 and Schultz et al. 1997 . For a survey of the theory and experimental support, see Glimcher 2011 .

As discussed in the chapter, neural architecture has been a model for the development of deep-learning architectures. For more on this, see Cichy et al. 2016 , Rajalingham et al. 2018 , and Ullman 2019 . There is a special issue of *Current Opinion in Neurobiology* on machine learning, big data, and neuroscience, with an overview in the editorial introduction (Pillow and Sahani 2019).

Chapter 13

Exploring Mindreading



Overview

13.1 Pretend Play and Metarepresentation

Leslie on Pretend Play and Metarepresentation

13.2 Metarepresentation, Autism, and Theory of Mind

Using the False-Belief Task to Study Mindreading

Interpreting the Results

Implicit and Explicit Understanding of False Belief

13.3 The Mindreading System

First Steps in Mindreading

From Dyadic to Triadic Interactions: Joint Visual Attention

TESS and TOMM

13.4 Mindreading as Simulation

Standard Simulationism

Radical Simulationism

13.5 The Cognitive Neuroscience of Mindreading

Neuroimaging Evidence for a Dedicated Theory-of-Mind System?

Neuroscientific Evidence for Simulation in Low-Level Mindreading?
Neuroscientific Evidence for Simulation in High-Level Mindreading?

Overview

This chapter introduces mindreading. This is a very general label for the skills and abilities that allow us to make sense of other people and to coordinate our behavior with theirs. Normal, encultured, nonautistic, and non－brain-damaged human beings are capable of an impressive degree of social coordination. We navigate the social world with the same skill and dexterity that we manifest in navigating the physical world. We can, almost effortlessly, work out what other people are trying to do and accommodate our behavior to suit theirs. Our mindreading skills are fundamental to social understanding and social coordination.

The dominant model of mindreading in cognitive science emerged from studies of pretending in young children, which are described in Section 13.1 . Section 13.2 builds up from pretending to mindreading, focusing on the *false-belief task* . Section 13.3 introduces a model of the entire mindreading system developed by developmental psychologist and autism specialist Simon Baron-Cohen. In Section 13.4 , we turn to a very different way of thinking about mindreading. According to the *simulation theory* , mindreading is achieved by running our " ordinary " information-processing systems offline in order to simulate how other people will solve a particular problem or react to a particular situation. Finally, in Section 13.5 , we turn to

cognitive neuroscience, exploring how the techniques and technologies discussed in Chapter 9 have been used to study mindreading.

13.1 Pretend Play and Metarepresentation

The emergence of pretend play is a major milestone in cognitive and social development, and many cognitive scientists have thought that it is a key precursor of mindreading skills.

Children start to engage in pretend play at a very young age, some as early as thirteen months. Normal infants are capable of engaging in fairly sophisticated types of pretend play by the end of their second year. The most basic type is *self-directed* — with the infant pretending to carry out some familiar activity (e.g., pretending to drink from an empty cup or to eat from a spoon with nothing on it). The next stage is *other-directed* , with the infant pretending that some object has properties it doesn ′ t have (e.g., pretending that a toy vehicle makes a sound or that a doll is saying something). More sophisticated is *object substitution* . This is when the infant pretends that some object is a different object and acts accordingly — pretends that a banana is a telephone, for example, and talks into it.

A common denominator in all instances of pretend play is that the infant represents objects and properties not perceptible in the object that is the focus of the pretense. This is the aspect of pretend play thought to lay the foundation for the emergence of mindreading abilities later in development. To see why, we need to look at the developmental scientist Alan Leslie ′ s influential model of pretend play.

Leslie on Pretend Play and Metarepresentation

Leslie observes that we cannot explain what is going on in pretend play simply with reference to infants ′ ordinary, nonpretend representations of objects. The infant can ′ t be somehow coordinating her ordinary banana representation and her ordinary telephone representation. After all, the banana is a banana, not a telephone. At the same time, though, the infant must recognize that although she is pretending that it is a telephone, what she has in front of her is really a banana — and representing the banana as a telephone requires representing it as having the properties that telephones standardly have.

Putting these ideas together, Leslie concluded that although representations in pretend play have their usual meaning, they cannot actually be functioning in a normal way. They are somehow " quarantined " from ordinary representations. Without this sort of quarantining, the infant ′ s representations of the world would be completely chaotic. One and the same cup would be both empty and contain water, for example. The key problem is to explain how the quarantining takes place.

Leslie ′ s explanation rests on a very basic parallel between how representations function in pretend play and how they function when we are representing other people ′ s mental states in mindreading. When we represent what other people believe or desire, we do so with representations that are also quarantined from the rest of our thinking about the world.

When I say, " Sarah believes that the world is flat, " I am asserting something about Sarah, namely, that she believes that the world is flat. But I am certainly not saying that the world is flat. If I were to utter the words " the world is flat " on their own, then I would standardly be making an assertion about the world. But when those very same words come prefixed

by the phrase " Sarah believes that . . . " they function very differently. They are no longer being used to talk about the world. I am using them to talk about Sarah ' s state of mind. They have become *decoupled* from their usual function.

Philosophers and psychologists typically describe this as a case of *metarepresentation* . Metarepresentation occurs when a representation is used to represent another representation rather than to represent the world. The fact that there is metarepresentation going on changes how words and mental representations behave. They no longer refer directly to the world. But they still have their basic meaning — if they lost their basic meaning, then they couldn ' t do the job of capturing how someone else represents the world.

The basic picture is summarized in Figure 13.1 . As the figure shows, ordinary (primary) representations can serve two functions. They can represent the world *directly* . This is the standard, or default, use. But they can also be used to metarepresent someone else ' s primary representations. This is what goes on when we engage in mindreading.

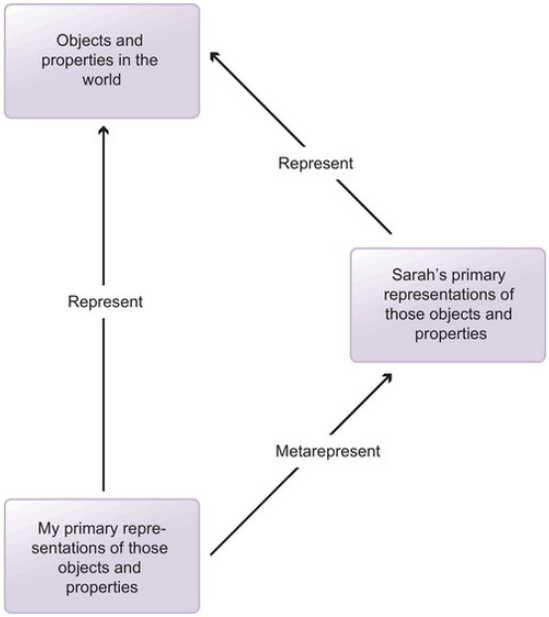


Figure 13.1 An example of metarepresentation. Metarepresentation involves second-order representations of representations. In this example, I am representing Sarah ′ s representations of certain objects and properties in the world.

Leslie ′ s basic idea is that pretend play involves a metarepresentational decoupling of ordinary representations from their usual functions. In fact, Leslie argues, the mechanism that decouples primary representations from their usual functions is exactly the same in pretend play and in mindreading. The structure of Leslie ′ s model is outlined in Figure 13.2 .

Exercise 13.1 Explain in your own words how metarepresentation differs from ordinary representation. Do you think that the distinction is a useful tool for understanding pretend play?

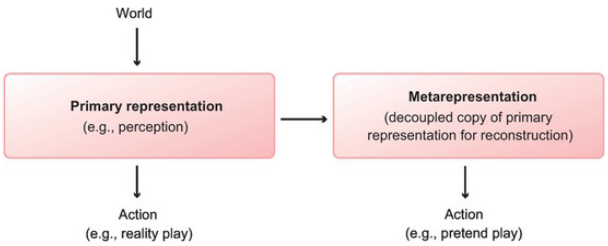


Figure 13.2 The general outlines of Leslie ' s model of pretend play.

(Adapted from Leslie 1987 .)

This is important because Leslie argues that the foundations for the mindreading system are laid during the second year of infancy, when infants acquire the basic machinery of decoupling and metarepresentation. We turn now to explaining how he thinks the journey from pretend play to mindreading takes place.

13.2 Metarepresentation, Autism, and Theory of Mind

In thinking about mindreading, Leslie paid careful attention to studies of children with autism. Autism is a developmental disorder thought to be genetic in origin, although its genetic basis is not very well understood. It typically emerges in toddlers, and the symptoms are often detectable before the age of two.

One feature of autism that particularly sparked Leslie ' s attention is that autistic children have well-documented problems with pretend play, in comparison both with much younger normal children and with children with intellectual disabilities of the same age. In fact, the phenomenon is so widespread in autism that it has become a standard diagnostic tool. Parents

are often first alerted to autism in their children by their apparent inability to engage in pretense and make-believe — and by the child ' s inability to understand what other people are up to when they try to incorporate the child into pretend play. And one of the first questions that clinicians ask when parents suspect that their child has autism is whether the child engages in pretend play.

Moreover, children with autism typically have deficits in social understanding, social coordination, and communication — in all of which mindreading plays an important part. But these social and communicative problems are not typically accompanied by general cognitive impairments. Autistic subjects can have very high IQs, for example.

In 1985, Leslie was one of the authors of a very influential paper arguing that children with autism had a very specific mindreading deficit — the other two authors were Simon Baron-Cohen and Uta Frith.

Using the False-Belief Task to Study Mindreading

Baron-Cohen, Leslie, and Frith studied three populations of children. The first group consisted of children with autism, aged between six and sixteen (with a mean of just under twelve). The second group of children had Down syndrome, which is a chromosomal disorder usually accompanied by mental disability, often severe. These varied from seven to seventeen years old (with a mean of ten). The third group (the control group) consisted of children with no cognitive or social disorders, aged from 3.5 to 6, with a mean of 4.5.

Baron-Cohen and colleagues used a famous experimental paradigm known as the *false-belief test* , first developed by developmental psychologists Heinz Wimmer and Joseph Perner in an article published in

1983. There are many different versions of the test, but they all explore whether young children understand that someone might have mistaken beliefs about the world. There is a very basic contrast between belief, on the one hand, and knowledge, say, on the other. There is no way in which I can *know* that some state of affairs holds without that state of affairs actually holding. Knowledge is an example of what philosophers sometimes call *factive* states.

> **Exercise 13.2** Can you give examples of other mental states that are factive in this sense?

In contrast, beliefs are not factive. I cannot have false knowledge, but I can (all too easily) have false beliefs. And anyone who understands what belief is needs to understand that it is possible to have false beliefs.

So, if a young child does not understand the possibility that someone might have false beliefs about the world, then they cannot understand what is involved in believing something. They cannot possess the concept of belief. And this, in turn, tells us something about their mindreading skills because children who do not understand the concept of belief are lacking a fundamental component of the mindreading toolkit.

Baron-Cohen et al.'s variant of Wimmer and Perner's original false-belief test is depicted in Figure 13.3. The child is seated in front of an experimenter, who has two puppets, Sally and Anne. Between child and experimenter is a table with a basket and box. In front of the child, Sally places a marble in the basket and then leaves the room. While she is away, Anne transfers the marble from the basket to the box. Sally then returns. The

experimenter asks the child: " Where will Sally look for her marble? " (or, in some versions of the test, " Where does Sally think the marble is? ").

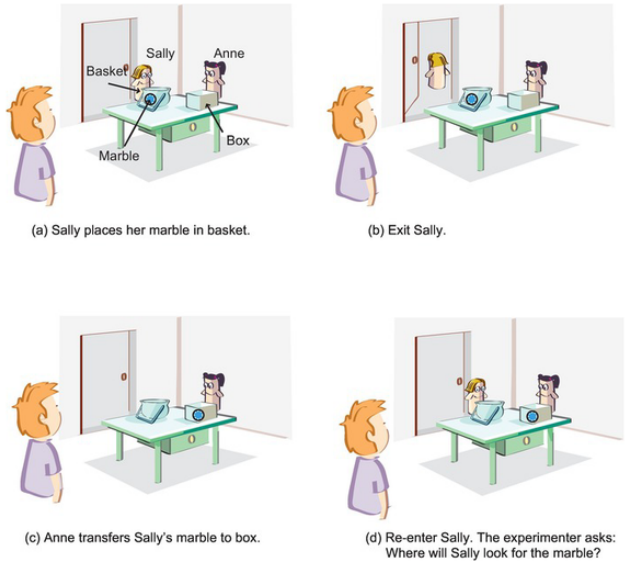


Figure 13.3 The task used by Baron-Cohen et al. to test for children ' s understanding of false belief.

(Adapted from Baron-Cohen et al. 1985 .)

The point of the experiment is that although the child saw the marble being moved, Sally did not. So, if the child has a clear grip on the concept of belief and understands that it is possible to have false beliefs, then she will answer that Sally will look in the basket because Sally has seen nothing to change her belief that the marble is in the basket. If, on the other hand, the child fails to understand the possibility of false belief, then she will answer that Sally will look for the marble where it in fact is, namely, in the box.

Exercise 13.3 Explain in your own words the logic behind the false-belief task. Do you think it succeeds in testing a young child ' s understanding of false belief?

Interpreting the Results

The main question that the experimenters asked was the obvious one, which they called the Belief Question: " Where will Sally look for her marble? " But they also wanted to make sure that all the children understood what was going on. So, they checked that each child knew which doll was which and asked two further questions:

"Where was the marble in the beginning?" (the Memory Question)
"Where is the marble really?" (the Reality Question)

Exercise 13.4 Explain in your own words the purpose of asking these two extra questions.

Baron-Cohen and colleagues found that all the children understood the experimental scenario. None of them failed either the Memory Question or the Reality Question. But there was a very significant difference in how the three groups fared with the Belief Question. Both the Down syndrome group and the normal group were overwhelmingly successful — with correct answers from 86 percent and 85 percent, respectively (despite the fact that the Down syndrome group had a mean verbal mental age of less than three). In striking contrast, the autistic group (with a mean verbal mental age of 5.5) performed extremely poorly. In fact, 80 percent of the children with autism failed the Belief Question.

The experimenters concluded that children with autism have a highly specific mindreading deficit. As they put it in the original paper, “ Our results strongly support the hypothesis that autistic children as a group fail to employ a theory of mind. We wish to explain this failure as an inability to represent mental states. As a result of this, the subjects with autism are unable to impute beliefs to others and are thus at a grave disadvantage when having to predict the behavior of other people ” (Baron-Cohen et al. 1985 : 43).

Notice the specific diagnosis of why the children with autism fail the false-belief task. It is described as a failure in the ability to represent mental states — in metarepresentation. The connection with Leslie ’ s theory of pretend play is illustrated in Figure 13.4 . The problems children with autism have with pretend play turn out to be very closely connected to their serious difficulties with the false-belief task — and with mindreading more generally. They have a common cause and a common explanation, namely, a deficit in metarepresentation.

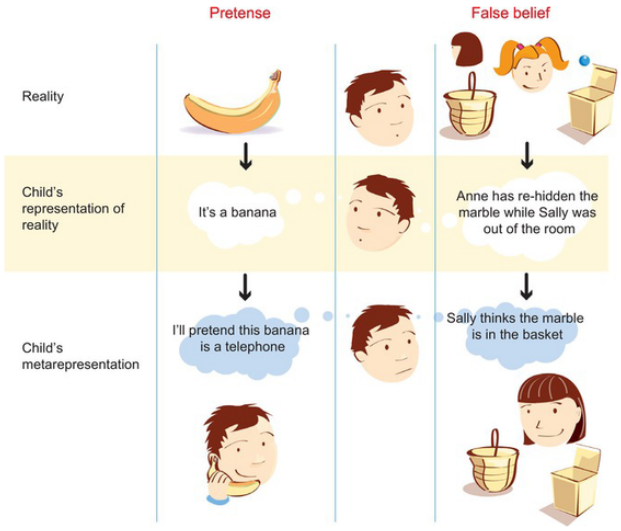


Figure 13.4 Illustration of the connection between pretend play and success on the false-belief task.

There remains a very important gap in the account, however. The problem is chronological. Pretend play emerges during the second year of life. But normal children do not typically pass the false-belief test until they are nearly four years old. Why is this? And what is the developmental progression that takes the normal child from pretend play to successful mindreading, as evidenced by success on the false-belief task? We will come back to these questions later. First, though, we need to consider some important experiments suggesting that children may be able to understand false beliefs significantly earlier than suggested by the standard false-belief task.

Implicit and Explicit Understanding of False Belief

The false-belief task originally proposed by Baron-Cohen and colleagues is a verbal task. Children are explicitly asked about where they think Sally will look or where they think the marble is. But it may be that these explicit questions introduce additional computational demands that muddy the waters. Perhaps young children fail the false-belief task because they cannot cope with these extra computational demands rather than because they do not understand false belief?

One way of exploring this possibility would be to develop a less demanding false-belief test. This was done by Kristine Onishi and Renée Baillargeon in a much-discussed set of experiments first published in 2005. Instead of explicitly asking children about how the characters they were observing would behave or what they believed, Onishi and Baillargeon used a violation-of-expectations paradigm that measured looking times. (Rather like the experiments on infant folk physics discussed in Chapter 11 .)

Fifteen-month-old infants were familiarized to an actor searching for a toy in one of two boxes (yellow and green, respectively). They were then presented with different conditions. In one condition, the toy was moved from one box to the other with the actor clearly watching. In a second condition, the toy was moved in the absence of the actor. After the toy was moved, the actor then looked for the toy in one of the two baskets.

Onishi and Baillargeon hypothesized that the length of time that the infants looked at each of the scenarios would indicate their degree of understanding of false belief. Consider the second scenario, where the toy is moved without the actor seeing. If the toy was moved without the actor observing, then the actor would presumably have a false belief about the toy ′s location, thinking it to be still in the green box when it is really in the yellow box. If infants understand this, then they will presumably expect the actor to search in the green box. And so they will look longer if the actor searches in the yellow box.

The experiments revealed a robust effect. Infants looked significantly longer when the actor searched in the yellow box than when the actor searched in the green box, even though the toy was really in the green box. Onishi and Baillargeon claim that the infants were surprised that the actor did not act on the basis of his (false) belief that the toy was still in the green box. So, they conclude, infants have an understanding of false belief much earlier than suggested by the traditional false-belief task.

However, the original, verbal false-belief experiments seem to be testing for a much more sophisticated cognitive ability than could be revealed by the Onishi and Baillargeon experiments. The earlier experiments target explicit conceptual abilities manifested in verbal responses and explicit reflection. Children are asked about what agents will do and what

they believe. What the experiments are getting at is mastery of the concept of belief, together with the complicated vocabulary and other baggage that goes with it.

In contrast, the Onishi and Baillargeon experiments are probing the nonverbal expectations that young children have about behavior and how behavior is affected by what an agent has and has not observed. It seems perfectly possible to have the right nonverbal expectations without being able to articulate them in the right sort of way to pass the false-belief test. In fact, considerable experimental evidence seems to suggest that this is what happens to most children between 1.5 and 4 years of age.

Perhaps the best way to look at the situation is this. The Onishi and Baillargeon experiments identify an *implicit* understanding of false belief, whereas the standard false-belief tasks are testing for an *explicit* understanding of false belief. An explicit understanding is verbally articulated and reflective, developed as part of high-level explanations of behavior in terms of beliefs and other mental states. An implicit understanding, in contrast, is predominantly practical, focused primarily on aligning one ′ s behavior with that of others and correctly predicting how others will behave as a function of what they have or have not seen.

Exercise 13.5 Explain in your own words the difference between implicit and explicit understandings of false belief.

13.3 The Mindreading System

We have explored some of the connections between mindreading and pretend play. The principal link between them, according to the model first proposed by Alan Leslie and developed by many others, is that both exploit metarepresentational skills — that is, a " decoupling " of representations from their usual functions.

However, young children start to engage in pretend play well before they are two years old, but it is not until around the age of four that they have a rich enough understanding of belief to pass the false-belief task. So, what happens in between?

First Steps in Mindreading

Developmental psychologist Simon Baron-Cohen was one of the coauthors of the 1985 paper that first drew the connection between autism and problems in mindreading. Since then, he has developed and fine-tuned a model of how mindreading emerges in infants and young children.

According to Baron-Cohen ' s model (depicted in Figure 13.5), the foundations of mindreading emerge in the earliest months of infant development, with the most basic mindreading skills in place by nine months of age. These skills are all essentially perceptual. The infant becomes perceptually sensitive to behavioral manifestations of psychological states.

- The *intentionality detector* (ID) allows the infant to identify purposeful movements. The ID codes movement as goal driven — it allows the infant to identify her mother ' s arm movement as reaching, for example. This allows the infant to distinguish animate, goal-driven entities from the other objects it encounters.

- The *eye-direction detector* (EDD) tracks other people ′ s eye movements. This is a good way of finding out the apparent goal of a purposeful movement because agents tend to keep their eyes on the target. The two mechanisms are highly complementary. There is little point in knowing that a movement is purposeful unless one has some idea what the goal is.
- The *emotion detector* (TED) is sensitive to the motivations and moods of the people with whom the individual is interacting — complete strangers, as well as caregivers and other family members. It allows infants to understand not just that agents make movements toward particular goals but also *why* those movements are being made and what sort of movements they are. Are they playful movements, for example, or protective ones?

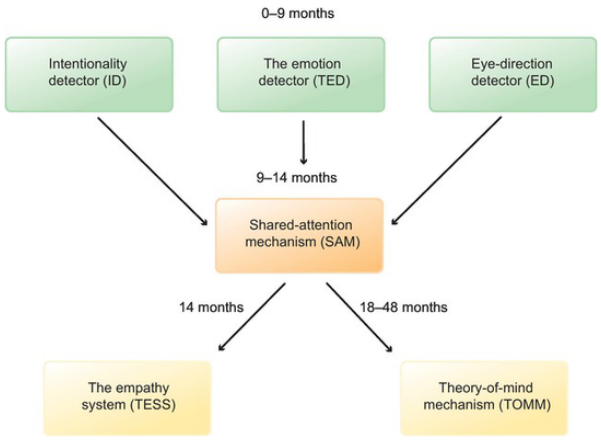


Figure 13.5 Baron-Cohen ′ s model of the mindreading system.

These three basic systems (ID, TED, and EDD) all involve relatively simple types of representation. TED, for example, involves " tagging " other agents with primitive representations of their moods (happy, sad, angry, frightened). EDD involves identifying a *dyadic* relation between an agent

and an object (*Dad sees the cup* , for example). Dyadic relations have two parts. The dyadic relation of seeing is a relation between an agent and an object. The ID also produces representations of dyadic relations between agents and objects. The dyadic relations here all involve intentional movements, such as reaching, following, or pushing .

From Dyadic to Triadic Interactions: Joint Visual Attention

A very important transformation takes place between the ages of nine and fourteen months. In the first nine months of life, infants can understand people and interact with them in certain very basic ways. They can also understand objects and manipulate them. For the very young infant, these are two separate activities. But starting at the age of nine months, the infant learns to combine them. Infants become capable of employing their interactions with people in their interactions with objects, and vice versa. This is illustrated in *joint visual attention* .

Joint visual attention occurs when infants look at objects (and take pleasure in looking at objects) because they see that another person is looking at that object — and because they see that the other person sees that they are looking at the object. It is a collaborative activity. Instead of just representing a dyadic relation between her mother and a cup, for example, the infant learns to represent *triadic* (or three-way) relations between herself, the mother, and the cup — as well as to initiate them with pointing and other gestures.

Joint visual attention becomes possible when the infant is able to embed representations — that is, to represent that an agent (whether herself or someone else) is representing someone else ′ s representation. This is very

different from the type of information processing involved in detecting eye direction or sensitivity to moods. It makes possible a whole range of coordinated social behaviors in which infants and their caregivers take enormous pleasure in collaborative games — games that involve and exploit an awareness of what others are doing and how they, too, are participating in the game.

This distinctive kind of information processing is carried out in what Baron-Cohen has termed the *shared-attention mechanism* (SAM). The connections with autism are very suggestive here too. It turns out that children with autism also have difficulties with joint attention — and that the more severe their social impairments are, the less able they are to engage in joint attention.

Shared attention is very important for language development because pointing to objects is a basic way of teaching children what words mean. In order for children to pick up on the cues that they are being given, they need to be able to understand that they and the person pointing are jointly attending to the very same thing.

TESS and TOMM

In Baron-Cohen ' s model, the SAM is a crucial foundation for the final two components of the mindreading system. The theory-of-mind mechanism (TOMM) incorporates an understanding of false belief — and of propositional attitudes in general. Earlier versions of Baron-Cohen ' s model contained only the TOMM after the SAM. Recently, however, he has added an additional component, which he calls TESS (for *the empathizing system*). For normal social development, it is not enough simply to be able to identify

other people ′ s emotional states and moods. The developing child needs to learn to respond appropriately to those emotional states and moods. This is where empathy comes in.

Psychosocial disorders such as psychopathy suggest that TOMM and TESS can come apart (and hence that there are distinct and separable mechanisms carrying out each task). Psychopaths are typically very good at working out what is going on in other people ′ s heads. The problem is that they tend not to care about what they find there — and in fact, they use their understanding to manipulate other people in ways that a normal person would find unacceptable. Diagnosing psychopathy is a very complex business, but psychiatrists typically put a lot of weight on basic failures of empathy — on failure to feel sympathy when someone else is in pain or obvious distress, for example.

TESS can only emerge if the basic capacity for shared attention is in place. In many ways, empathy is a matter of being able to put oneself in someone else ′ s position — to imagine what it would be like to be someone else and to find oneself in the situation that they find themselves in. Shared attention basically exploits the same ability; it is just being applied in a much more limited sphere. The child engaged in joint visual attention or collaborative play is able to adopt someone else ′ s visual perspective to represent how things look to someone else. As they do this more and more, they eventually bootstrap themselves into the ability to understand someone else ′ s emotional perspective on the world — to understand not just how a situation looks to someone but also how that situation affects them .

13.4 Mindreading as Simulation

Up to now, we have been considering mindreading primarily as a theoretical accomplishment — that is, as bringing a specialized body of knowledge (theory of mind) to bear in order to explain and predict the behavior of others.

This section explores an alternative to this theory-based approach. According to the *simulation theory* , there is no specialized theory-of-mind mechanism. Instead, theory-of-mind processing is carried out by the very same systems that are responsible for ordinary decision-making and for finding out about the world.

Here is the basic idea. Suppose that we have a reasonable sense of the beliefs and desires that would be appropriate to attribute to someone else in a particular situation. To find out how they will behave, we don ′ t use specialized knowledge about how mental states typically feed into behavior. Instead, we use our own decision-making processes to run a simulation of what would happen if we ourselves had those beliefs and desires. We use the output from that simulation process to predict how the person will behave.

Standard Simulationism

What I will term *standard simulationism* was originally proposed by developmental psychologist Paul Harris and subsequently developed by philosopher Alvin Goldman. It holds that simulation starts with the mindreader *explicitly* (although not necessarily consciously) forming judgments about how the other person represents the relevant situation and what they want to achieve in that situation. These judgments serve as pretend inputs to the ordinary decision-making system. This decision-

making system then runs offline to yield predictions as to how the other person will behave. The general model is illustrated in Figure 13.6 .

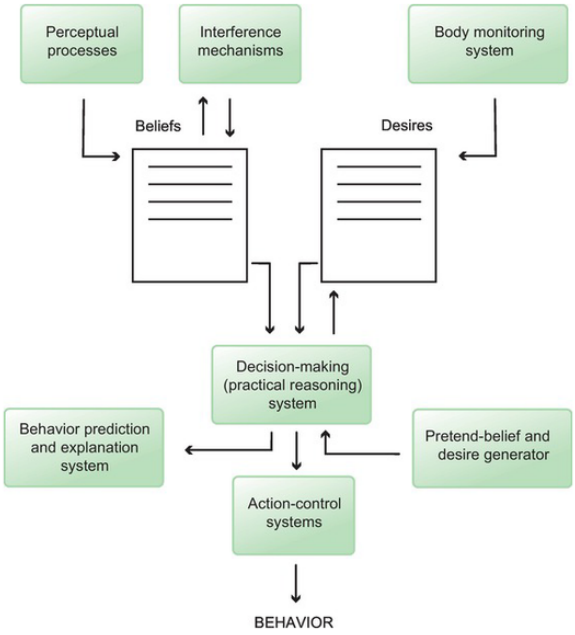


Figure 13.6 A schematic version of standard simulationism. Note that the ordinary decision-making system is being run offline with pretend inputs.

(Adapted from Nichols et al. 1996 .)

Where do these judgments come from? Goldman thinks that we work by analogy from our own beliefs and desires. We know which beliefs we tend to form in response to particular situations. And so, we assume that others will form the same beliefs, unless we have specific evidence to the contrary — for example, about how they have acted in the past, or about temperamental differences, or the different information that they might have. When we do have such additional evidence, we make adjustments by thinking about what we would do if we had those temperamental features or that extra information.

So, in Goldman ′ s model, knowledge of others rests on self-knowledge. And he thinks that we have a special mechanism for finding out about our own beliefs, desires, and other propositional attitudes — a self-monitoring mechanism that philosophers call *introspection* or *inner sense* .

Radical Simulationism

There is a second way of developing the basic simulationist idea — often called *radical simulationism* — developed primarily by philosophers Robert Gordon and Jane Heal. The intuitive idea here is that instead of coming explicitly to the view that the person whose behavior I am trying to predict has a certain belief (say, the belief that *p*), what I need to do is to imagine how the world would appear from her point of view.

Here is the difference. In standard simulationism, the process of simulation starts with beliefs about another person ′ s beliefs and desires. They are metarepresentations. According to radical simulationism, on the other hand, what the simulator is thinking about is the world rather than the person they are simulating. The simulator is thinking about the world *from the perspective of the person being simulated* rather than thinking about their beliefs, desires, and other psychological states.

Radical simulationism proposes mindreading without metarepresentation. It is world directed rather than mind directed. And as a result, it gives a very different account of what is going wrong when children fail the false-belief test. For the radical simulationist, children who fail the false-belief test lack imaginative capacities. Their capacity to project themselves imaginatively into someone else ′ s position is not sufficiently

developed. They are not yet able to form beliefs *from a perspective other than their own* .

Exercise 13.6 We have now looked at four different ways of thinking about the false-belief task. Draw up a table indicating the four different proposals that have been made for explaining what it is that the false-belief task is testing for.

13.5 The Cognitive Neuroscience of Mindreading

The cognitive neuroscience of mindreading has become a hot topic in recent years, and we can only scratch the surface here. I will focus on three questions that emerge from our earlier discussion.

Several of the models considered hold that the mind contains a dedicated TOMM — an information-processing system dedicated to reasoning about other people ′ s beliefs, desires, and other propositional attitudes. So, a natural question to ask is as follows:

Question 1: Is there any evidence at the neural level for the existence of a TOMM?

We also looked at the alternative proposal that mindreading is an exercise in simulation. It does not exploit systems specialized for mindreading. Instead, the information processing in mindreading is carried

out by cognitive systems that also do other things. We can call these *co-opted systems.* So, we can ask whether there is any evidence at the neural level for this way of thinking about mindreading.

There are really two different questions here. We can make a distinction between low-level mindreading and high-level mindreading. Low-level mindreading involves detecting emotions, identifying goal-driven actions, sensitivity to eye gaze, and so on. High-level mindreading involves identifying and reasoning about beliefs, desires, and other psychological states. This gives us two further questions:

> *Question 2:* Is there evidence at the neural level that *low-level mindreading* is a process of simulation involving co-opted systems?

> *Question 3:* Is there evidence at the neural level that *high-level mindreading* is a process of simulation involving co-opted systems?

Neuroimaging Evidence for a Dedicated Theory-of-Mind System?

To investigate whether there is a dedicated TOMM, experimenters have looked for the following:

1. Brain regions showing increased activity in response to information-processing tasks that require the subject to attribute beliefs

2. Brain regions where these increased activation levels are specific to tasks involving belief attribution — as opposed, for example, to

reflecting demands on general reasoning or the fact that people (rather than inanimate objects) are involved

With respect to (1), a candidate TOMM region should show increased activation both for false-belief tasks and for true-belief tasks. To establish that (2) holds, experimenters need to make sure that they have controlled for domain-general processes (such as language or working memory).

Neuroimaging studies have identified a number of brain regions with increased activation in tasks that require reasoning about beliefs. Most have used versions of the false-belief test, but not all. Cognitive psychologist Vinod Goel, for example, ran a series of studies in which he asked subjects to decide whether Christopher Columbus would have been able to work out the function of an object from a picture — the idea being that this task requires subjects to reason about the sort of beliefs that a fifteenth-century explorer would have been likely to have. Other studies had subjects read a short story and then answer questions about it. Some of the questions required making inferences about the beliefs of characters in the story, and others did not.

Studies such as these have identified a number of brain regions as potentially forming part of a dedicated theory-of-mind system. These include the following regions (working more or less from front to back):

- Medial prefrontal cortex
- Anterior cingulate cortex
- Orbitofrontal cortex
- Temporal pole
- Broca ′ s area

- Anterior superior temporal sulcus
- Fusiform gyrus
- Temporoparietal junction
- Posterior superior temporal sulcus

This is a long list, and as Figure 13.7 shows, these regions collectively cover a large area of the brain.

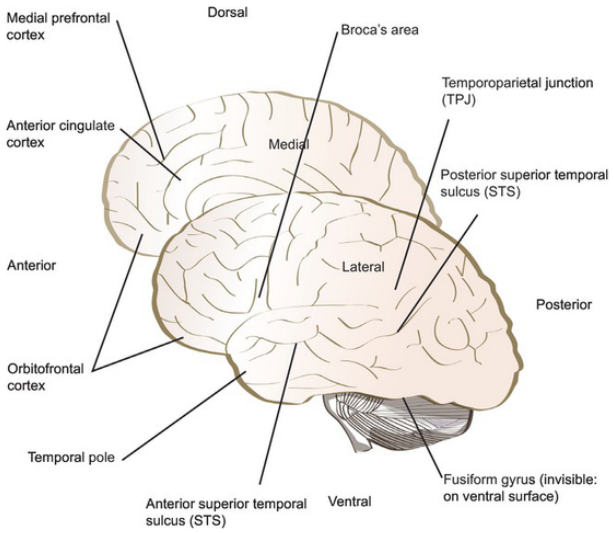


Figure 13.7 Schematic representation of brain regions associated with the attribution of mental states.

(Adapted from Saxe et al. 2004 .)

A number of the areas are thought to be specialized for other information-processing functions. Broca ′ s area, for example, is widely held to be involved in aspects of language processing, whereas the fusiform gyrus includes the fusiform face area (which has been hypothesized as a dedicated face-processing system). Language and face recognition are both involved in mindreading. So, to narrow the list down, we need to see which (if any) of these areas satisfy (1) and (2) in the previous list.

The first stage, corresponding to (1), is to check whether particular neural regions show activation both in false-belief and true-belief conditions. Many neuroimaging studies follow the developmental studies in focusing only on false-belief conditions. But if we are looking for a neural substrate for belief reasoning, we need to consider true-belief conditions as well as false ones — after all, perhaps some of the activation in the false-belief condition is due to the falsity of the belief attributed rather than to its being a belief.

Rebecca Saxe and Nancy Kanwisher carried out a set of false-belief experiments with a true-belief condition as a control (more on these experiments later in the chapter). They found three brain regions where both true- and false-belief attribution tasks elicited activation in the very same voxels. (Recall that a voxel is a *volumetric pixel* representing a small volume within the brain.) These regions are as follows:

- Medial prefrontal cortex (MPFC)
- Superior temporal sulcus (STS)
- Temporoparietal junction (TPJ)

Applying (1), then, significantly narrows down the field. What happens when we apply (2)?

To apply (2), we need to control for other types of domain-general information processing that might generate activation. Saxe and Kanwisher introduced two control conditions, based on their analysis of what is required in order to succeed on tasks involving belief attribution.

Their first observation is that when we attribute beliefs to other people, we are effectively identifying hidden causes. This is because we typically attribute beliefs when we are trying to explain or predict behavior, and we

cannot do so in terms of what is immediately observable. So, to make sure that activation really does reflect domain-specific theory-of-mind reasoning, we need to rule out the possibility that what is going on is really just domain-general reasoning about hidden causes.

To do this, Saxe and Kanwisher developed a set of stories depending on nonpsychological hidden causes. Here are two:

- The beautiful ice sculpture received first prize in the contest. It was very intricate. Unfortunately, the temperatures that night hit a record high for January. By dawn, there was no sculpture.
- The night was warm and dry. There had not been a cloud anywhere for days. The moisture was certainly not from rain. And yet, in the early morning, the long grasses were dripping with cool water.

Call this the *hidden-causes condition* .

Saxe and Kanwisher also wanted to rule out the possibility that activation is due to general reasoning about false representations — as opposed to false *beliefs*. There is nothing psychological about a false representation such as a misleading map, for example. To rule out the possibility that the neural areas active in belief attribution are specialized for information processing having to do with representations in general rather than theory of mind, Saxe and Kanwisher used a version of the *false-photograph task* originally proposed by developmental psychologist Debbie Zaitchik.

As before, the subject is presented with a story in which Sally places a marble in the basket. A photograph is taken of the contents of the basket and placed facedown. After the photograph is taken, Anne moves the marble from the basket to the box. Everything is exactly as in the false-belief task

— except that the subjects are asked where the object appears in the photograph. A subject who does not understand the possibility of false representations will think that the object ′ s location in the photograph will be where it really is — and so will answer that the photograph will depict the marble as being in the box.

Exercise 13.7 Explain and assess the reasoning behind the false-photograph task .

Saxe and Kanwisher found that there was significant activation in the three regions identified earlier (MPFC, STS, and TPJ) in the belief-attribution condition, but not in the false-representation or hidden-causes conditions.

Still, all this shows is that there are brain regions specialized for processing information about mental states such as belief. These experiments do not tell us much, if anything, about what is going in those regions.

Neuroscientific Evidence for Simulation in Low-Level Mindreading?

Simulation theorists claim that mindreading is carried out by what they call *co-opted mechanisms*. These are information-processing systems that normally serve another function and that are then recruited to help make sense of the social world. Numerous experiments have been interpreted by simulation theorists as showing that co-opted mechanisms play a fundamental role in mindreading.

One very basic form of mindreading is the ability to read emotions from perceptible bodily states. Facial expressions are the most obvious example, but tone and timbre of voice are often good guides to emotions, as are global features of posture (the vitality and energy of someone ' s movements, for example). Young children start to develop their skills in this form of mindreading at a very early age. It is an automatic and unconscious process for normal people — fundamental to our interactions with others. In Baron-Cohen ' s model, it is the job of a dedicated component: the emotion detector.

In the simulationist view, the emotion detector is a co-opted mechanism (or set of mechanisms), so there is a single set of emotion mechanisms in play both when agents are experiencing emotional states and when they detect emotions in others. Is there any evidence that this is so? Some suggestive results have come from the study of patients with brain damage.

Many studies have found that a region of the temporal lobe known as the *amygdala* plays an important role in mediating fear. The experience of disgust, in contrast, is much more closely correlated with activity in the *insula* , which lies in the lateral sulcus, separating the temporal lobe from the parietal lobe. (Both the amygdala and the insula form part of the *limbic system.*)

Simulationism predicts that the amygdala will be active both when someone is undergoing fear and when they identify that fear in others. And conversely, it would expect damage to the amygdala to result in a patient having problems *both* with the experience of fear and with identifying fear in others. In other words, damage to brain regions that play a significant role in mediating particular emotions will result in *paired deficits* — in problems

with experiencing the relevant emotion and in identifying it in others. There is evidence of paired deficits for several different emotional states:

- *Fear:* Ralph Adolphs and his colleagues have studied a number of patients with damage to the amygdala. The patient S.M., for example, had her amygdala destroyed on both sides of the brain by Urbach – Wiethe disease. She is, quite literally, fearless — although she knows what fear is, she does not experience it. She is also significantly impaired on tests that require identifying fear on the basis of facial expression. Patients with psychopathology are known to have both smaller amygdalas than normal subjects and reduced capacities for experiencing fear. It turns out that they are also much less effective than normal controls at identifying fear in others.
- *Anger:* The neurotransmitter *dopamine* is thought to play an important role in the experience of anger. Experiments on rats, for example, have shown that levels of aggression can be directly manipulated by raising/lowering the rat ′s dopamine levels. In humans, dopamine production can be temporarily blocked with a drug called *sulpiride.* Experiments have shown that subjects whose dopamine levels have been lowered in this way are significantly worse than controls at recognizing anger from facial expression — but they do not have problems with other emotions.
- *Disgust:* The brain area most associated with the experience of disgust is the insula. Neuroimaging studies have shown that this area is also activated when subjects observe facial expressions of disgust. This result is confirmed by studies of patients with brain damage. N.K., a much-studied patient suffering from damage to the insula and

basal ganglia, has severe problems in both experiencing disgust and recognizing it in others. He performs no differently from controls, however, with regard to other basic emotions (such as surprise and fear).

Simulationists have also found evidence for coopted mechanisms in some much-publicized experiments on "mirror neurons." We looked briefly at mirror neurons in Section 9.2 as an example of what we can learn from recording electrical activity in single neurons. (This would be a good moment to look back at Figure 9.5 to see mirror neurons in action.)

Mirror neurons were first discovered in macaque monkeys by an Italian research group led by Giacomo Rizzolatti in the mid-1990s. Rizzolatti and his colleagues were recording the responses of neurons that showed selective activation when the monkey made certain hand movements (such as reaching for a piece of food) when they noticed, completely by chance, that the same neurons fired when the monkey saw an experimenter making the same movement.

In monkeys, the mirror-neuron system is located in area F5 in the ventral premotor cortex, as well as in the inferior parietal lobe. No mirror neurons have ever been directly detected in humans — not surprisingly because it is not usually possible to make single-cell recordings in humans. The evidence for mirror neurons in humans comes primarily from functional magnetic resonance imaging (fMRI) studies. Studies have found a brain system that appears to have the basic "mirroring" feature — that is, its elements show activation both when the subject performs certain actions and when others are observed making that action. Researchers have dubbed this system the *mirror-neuron system*. The *mirror-neuron system* is illustrated in Figure 13.8 and described in the accompanying caption.

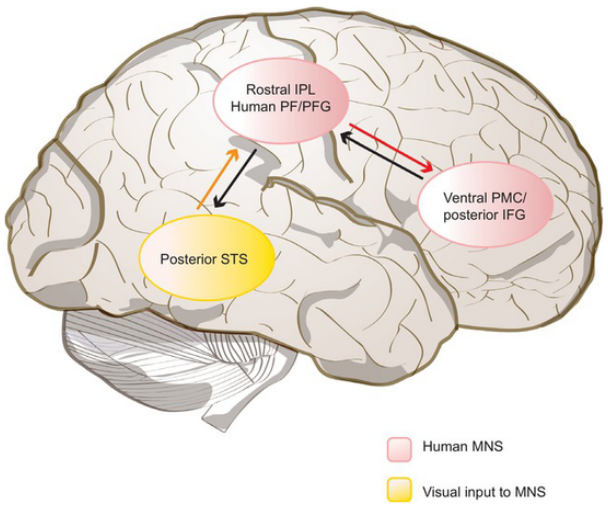


Figure 13.8 Schematic overview of the frontoparietal mirror-neuron system (MNS) (*pink*) and its main visual input (*yellow*) in the human brain. An anterior area with mirror-neuron properties is located in the inferior frontal cortex, encompassing the posterior inferior frontal gyrus (IFG) and adjacent ventral premotor cortex (PMC). A posterior area with mirror-neuron properties is located in the rostral part of the inferior parietal lobule (IPL) and can be considered the human homolog of area PF/PFG in the macaque. The main visual input to the MNS originates from the posterior sector of the superior temporal sulcus (STS). Together, these three areas form a " core circuit " for imitation. The visual input from the STS to the MNS is represented by an orange arrow. The red arrow represents the information flow from the parietal MNS, which is mostly concerned with the motoric description of the action, to the frontal MNS, which is more concerned with the goal of the action. The black arrows represent reference copies of motor predictions of imitative motor plans and the visual description of the observed action.

(Adapted from Iacoboni and Dapretto 2006.)

Some cognitive scientists have suggested that the mirror-neuron system functions as an *empathy* system. It allows people to *resonate* with the

psychological states of other people. So, for example, studies have shown that areas in the mirror-neuron system are activated both when the subjects feel pain and when they observe a loved one undergoing a painful stimulus. In terms of the models that we have been using, this would mean that the mirror-neuron system could serve as a neural substrate both for the TED (the emotion detector system) and the TESS (the empathy system). And as the caption to Figure 13.8 brings out, it is also thought that the mirror-neuron system is part of what makes imitation possible.

A possible application for mirror neurons is in detecting basic forms of goal-driven action — Baron-Cohen ′s intentionality detector. There is some evidence that mirror neurons are sensitive to goals (rather than simply to bodily movements). A study published in 2001 by Alessandra Umilta and colleagues showed that mirror neurons fire even when the monkey cannot see the final stages of the action. They used a screen to hide the experimenter ′ s hand when it actually grasped the object and found that about 50 percent of the mirror neurons usually tuned to grasping actions were activated even in the absence of the usual visual cues for grasping. It seems that mirror neurons are sensitive to fairly abstract properties of movements — to the fact that they are goal directed rather than simply to their physical and observable characteristics.

In any event, mirror neurons in monkeys are direct examples, at the most basic neural level, of mechanisms that show the *dual-purpose* structure at the heart of the simulationist approach to mindreading. And much of the evidence that has been produced in support of the existence of a mirror-neuron system points to the existence of brain regions that serve both first-person and third-person roles. They are active both when the subject performs certain actions and/or undergoes experiences of a certain type and

when others are observed performing those actions and/or undergoing those experiences .

Neuroscientific Evidence for Simulation in High-Level Mindreading?

There is less direct evidence for simulation in high-level mindreading than in the lower-level processes that we have just been discussing. Nonetheless, there are some suggestive results.

For standard simulationists, the process of simulation requires some form of *inference by analogy.* So, if standard simulation is a correct way of thinking about mindreading, then it should be both a first-person and a third-person process. The basic engine of simulation is the simulator running her own decision-making processes offline and identifying her own mental states. The results of this first-person simulation are then applied to the person being simulated. The prediction from standard simulation, therefore, is that regions of the brain specialized for what is sometimes called *self-reflection* (i.e., identifying one ′ s own psychological attributes, abilities, and character traits) will be active during tasks that require mindreading.

There is some evidence bearing this prediction out. Some studies have shown that self-reflection tasks elicit activation in an area of the brain thought to be involved in high-level mindreading — the medial prefrontal cortex (MPFC — illustrated in Figure 13.7). So, for example, William Kelly and collaborators in 2002 presented subjects with various written adjectives and asked some questions about them. These questions were either perceptual (‟ Is this adjective written in italics? ″), self-directed (‟ Does this adjective describe you? ″), or other directed (‟ Does this adjective

describe the president? ″). The self-directed questions consistently generated greater activation in the MPFC.

Further support comes from a study published in 2005 by Jason Mitchell, Mazharin Banaji, and Neil Macrae. They scanned subjects while they were presented with photographs of other people and asked questions about them. Some questions required mindreading (‘‘ How pleased is this person to have their photograph taken? ″), whereas others did not (‘‘ How symmetrical is this person ′ s face? ″). After a short delay, the subjects were presented with the photographs again and asked how similar they thought the other person was to themselves. This question is important for simulation theorists because simulation is likely to work best for people who one thinks are similar to oneself.

These experiments produced two significant results. First, they provided further evidence that the MPFC is important in high-level mindreading — the MPFC showed much higher activation levels on the mindreading version of the task than on the other version. Moreover, there was a significant correlation between activation in the MPFC while the subjects were answering the mindreading questions and the degree of similarity that subjects subsequently observed between themselves and the person in the photograph. The greater the perceived similarity with the person in the photograph, the higher the level of activation in the subject ′ s MPFC.

It is far too early to draw any definite conclusions about the cognitive neuroscience of mindreading. But even this brief review illustrates two important points from earlier chapters:

- The cognitive neuroscience of mindreading involves careful calibration of results from different technologies. So, for example, experimenters working through the potential implications of mirror neurons for mindreading in humans have calibrated single-neuron studies in monkeys with functional neuroimaging in humans.
- Participants at all levels in the debate clearly recognize that techniques from neuroscience have a crucial role to play in testing, confirming, and developing cognitive models .

Summary

This chapter has explored theoretical models of mindreading — the suite of skills and abilities that allows us to make sense of other people and navigate the social world. The chapter began by reviewing Leslie ′ s theory that mindreading exploits the basic abilities that are also deployed in pretend play. These are abilities for metarepresentation — for representing representations. We looked at a famous set of experiments using the false-belief task that seemed to show that children with autism (who are known to be deficient in pretend play) are also impaired in tasks involving reasoning about other people ′ s beliefs. Mindreading is a complex phenomenon, and we looked at a model of mindreading that sees it as made up of six distinct components, emerging at different stages in cognitive development.

We then explored an alternative model of mindreading. According to simulationists, there is no dedicated mindreading system. Instead, mindreading is carried out by our “ ordinary ” cognitive systems running

offline with pretend inputs. We looked at two different versions of simulationism — standard simulationism and radical simulationism.

Finally, we reviewed a range of evidence from cognitive neuroscience. Some neuroimaging experiments reveal areas with increased activation during mindreading tasks. At the same time, there is evidence for the simulationist hypothesis that mindreading involves co-opting ordinary decision-making systems, both from research on mirror neurons in monkeys and also from imaging studies showing that areas specialized for self-reflection are active in mindreading tasks.

Checklists

Alan Leslie ' s model of mindreading in young children is based on an analogy with the information processing involved in pretend play.

1. The emergence of pretend play in the second year of life is a major milestone in cognitive and social development.

2. In pretend play, some of an infant ' s *primary representations* of the world and other people become " decoupled " from their usual functions while preserving their ordinary meaning.

3. Leslie thinks that primary representations function in the same way in pretend play as in mindreading. Both pretend play and mindreading exploit *metarepresentation* .

4. Children with autism have significant problems both with mindreading and with pretend play.

5. The false-belief task (developed by Heinz Wimmer and Joseph Perner) is a standard test of mindreading abilities in children. It tests whether children are able to abstract away from their own knowledge to understand that someone else can have different (and mistaken) beliefs about the world.

High-level mindreading involves attributing *propositional attitudes* (such as beliefs and desires) to other people. But high-level mindreading depends on a complex system of lower-level mechanisms — as in Simon Baron-Cohen ′s model of the overall mindreading system.

1. The *intentionality detector* is responsible for perceptual sensitivity to purposeful movements.

2. The *eye-direction detector* makes it easier to identify the goals of purposeful movements and to see where other people ′s attention is focused.

3. The *emotion detector* gives a basic sensitivity to emotions and moods, as revealed in facial expressions, tone of voice, and so forth.

4. The *shared-attention mechanism* makes possible a range of coordinated social behaviors and collaborative activities.

5. The *empathizing system* is responsible for affective responses to other people ′s moods and emotions (as opposed to simply identifying them).

The simulationist approach to mindreading denies that there is a dedicated theory-of-mind system responsible for identifying and

reasoning about other people ′ s beliefs, desires, and other propositional attitudes.

1. Simulationists think that mindreading is carried out by " ordinary " information-processing systems that are co-opted for mindreading. We use our own minds as a model of someone else ′ s mind.

2. According to standard simulationism, we predict other people ′ s behavior, for example, by running our decision-making processes offline, with pretend beliefs and desires as inputs.

3. Radical simulationists hold that mindreading involves representing the world from another person ′ s perspective, without any representation of their psychological states.

Cognitive neuroscientists have used a range of techniques, including single-neuron recording and functional neuroimaging, in order to test and refine cognitive models of mindreading.

1. Neuroimaging studies have identified a number of brain areas with increased activation in mindreading tasks — for example, the medial prefrontal cortex, the superior temporal sulcus, and the inferior parietal lobule. This is consistent with the claim that there is a dedicated theory-of-mind system.

2. There is evidence that co-opted mechanisms are used in low-level mindreading (as predicted by the simulation theory). Areas active during the experience of basic emotions, such as fear, disgust, and anger, are also active when those emotions are identified in others.

3. Mirror neurons in area F5 of the macaque brain respond both when the monkey performs an action and when the monkey observes an

experimenter or conspecific perform that action. Researchers have hypothesized a mirror-neuron system in the human brain. This may play an important role in understanding goal-directed action.

4. There is evidence consistent with simulation-driven processing in high-level mindreading. Experiments have shown that areas specialized for self-reflection are also implicated in mindreading (as predicted by standard simulationism).

Further Reading

Leslie first presented his metarepresentational theory of pretend play and mindreading in Leslie 1987 . For further developments, see Leslie and Polizzi 1998 , Leslie et al. 2004 , Leslie et al. 2005, Friedman and Leslie 2007, and Freiedman et al. 2010. For more general research on pretend play, see Weisberg 2015 , Lillard 2017 , Ma and Lillard 2017 , and Goldstein and Lerner 2018 .

The false-belief task discussed in the text was first presented in Wimmer and Perner 1983 . For powerful criticisms of its claimed relevance to theory of mind, see Bloom and German 2000 . Perner ' s own theory of mindreading is presented in his book *Understanding the Representational Mind* (1993). Wellman 2018 gives an overview of more recent research. See Phillips et al. 2021 with accompanying peer commentary for the proposal that researchers on mindreading should focus on knowledge rather than false belief.

The claim that young infants can pass a version of the false-belief task was made in Onishi and Baillargeon 2005 . For discussion, see Perner and

Ruffman 2005 , Rakoczy 2012 , and Heyes 2014 . Poulin-Dubois et al. 2009 reviews studies on infant mindreading abilities. Numerous reviews discuss both implicit and explicit false-belief understanding (Baillargeon et al. 2010 ; Beate 2011 ; Low and Perner 2012 ; Luo and Baillargeon 2010 ; Perner and Roessler 2012 ; Trauble et al. 2010). For a philosophical perspective on this research, see Carruthers 2013 . Schuwerk et al. 2018 and Wenzel et al. 2019 report failures to replicate important studies on implicit mindreading. See also Barone et al. 2019 .

The idea that autism is essentially a disorder of mindreading was first presented in Baron-Cohen et al. 1985 . For a book-length discussion of autism as " mindblindness, " see Baron-Cohen 1995 . This interpretation of autism has been challenged — see, for example, Boucher 1996 . The papers in Baron-Cohen et al. 2000 discuss autism from the perspective of developmental psychology and cognitive neuroscience. For a survey, see Baron-Cohen 2009 . Fletcher-Watson et al. 2014 reviews studies of clinical interventions based on the mindreading theory of autism. Roeyers and Denurie 2010 and Mathersul et al. 2013 discuss alternatives to the false-belief task usable in adolescents and adults. Mundy 2018 reviews the neuroscience of joint attention in neurotypical subjects and those with autism.

Mindreading was one of the earliest fields to see sustained interactions and collaborations between philosophers and psychologists. Numerous influential early papers, including Heal 1986 and Gordon 1986 , are gathered in two anthologies edited by Davies and Stone (1995a , 1995b) and in Carruthers and Smith 1996 . All have useful introductions. Recent contributions to the dialog include Nicholson et al. 2019 and 2021 . Much of this debate focuses on comparing simulationist approaches to mindreading

(as presented in Section 12.5) with the more traditional approach discussed in earlier sections (what is often called the *theory model of mindreading*). Goldman 2006 is a book-length defense of simulationism, written by a philosopher but with extensive discussions of the empirical literature.

Studies on the cognitive neuroscience of mindreading include Apperly et al. 2004 , Samson et al. 2004 , Samson et al. 2005 , Saxe and Kanwisher 2005 , Saxe et al. 2004 , Tamir and Mitchell 2010, and Waytz and Mitchell 2011 . Reviews can be found in Adolphs 2009 , Saxe 2009 , Carrington and Bailey 2009 , Abu-Akel and Shamay-Tsoory 2011 , Frith and Frith 2012 , and Schurz and Perner 2015 . Claims about the modularity of mindreading are critically discussed in Apperly et al. 2005 . For skepticism about the false-photograph task, see Perner and Leekam 2008 .

Research into mirror neurons has been reported in many papers — see, for example, Rizzolatti et al. 2001 . The findings are presented for a general audience in Rizzolatti et al. 2006 (article) and Rizzolatti and Sinigaglia 2008 (book). See also Rizzolatti and Sinigaglia 2010 and Rizzolatti and Fogassi 2014 . Cook et al. 2014 presents an alternative perspective on mirror neurons, and Kilner and Lemon 2013 offers an independent assessment of the literature. Heyes and Catmur 2022 reviews recent work on mirror neurons, whereas Yates and Hobson 2020 discusses mirror neurons and autism (including the so-called *broken-mirror hypothesis*).

For more information on empirical findings about emotion recognition in patients with brain damage and normal controls, see Adolphs et al. 1994 , Phillips et al. 1997 , Adolphs and Tranel 2000 , and Wicker et al. 2003 . See also Croker and Macdonald 2005 and (for a review) Bornhofen and McDonald 2008 .

Chapter 14

Robotics: From GOFAI to Situated Cognition and Behavior-Based Robotics

Overview

14.1 GOFAI Robotics: SHAKEY

SHAKEY ′s Software I: Low-Level Activities and Intermediate-Level Actions

SHAKEY ′s Software II: Logic Programming in STRIPS and PLANEX

14.2 Situated Cognition and Biorobotics

The Challenge of Building a Situated Agent

Situated Cognition and Knowledge Representation

Biorobotics: Insects and Morphological Computation

14.3 From Subsumption Architectures to Behavior-Based Robotics

Subsumption Architectures: The Example of Allen

Behavior-Based Robotics: TOTO

Multiagent Programming: The Nerd Herd

Overview

This chapter focuses on robotics. We start with a classic example of GOFAI robotics — *GOFAI* stands for *good old-fashioned artificial intelligence* . The remainder of the chapter explores the situated-cognition movement in robotics. Like dynamical systems theorists (see Chapter 6), situated-cognition theorists are dissatisfied with traditional ways of thinking about information processing in cognitive science. They have developed a powerful toolkit of alternatives and used them to construct new types of robots.

Section 14.1 reviews one of the historic achievements of early robotics, which is also an excellent illustration of the physical symbol system hypothesis in action. This is SHAKEY, a mobile robot developed at the Artificial Intelligence Center at Stanford Research Institute (SRI). SHAKEY was designed to operate and perform simple tasks in a real, physical environment. The programs built into it permitted SHAKEY to plan ahead and to learn how to perform tasks better.

Section 14.2 brings out some of the complaints that situated-cognition theorists level at traditional GOFAI robotics and illustrates some of the engineering inspiration that these theorists have drawn from studying very simple cognitive systems, such as those of insects.

Section 14.3 explores how these theoretical ideas have been translated into particular robotic architectures, focusing on the subsumption architectures developed by Rodney Brooks, and looks at examples of what Maja Matarić has termed *behavior-based robotics* .

14.1 GOFAI Robotics: SHAKEY

This section looks at a pioneering robot developed in the late 1960s and early 1970s in the Artificial Intelligence Center at what was then called the Stanford Research Institute (now SRI International and no longer affiliated with Stanford University). This robot, affectionately called SHAKEY (because of its jerky movements), was the first robot able to move around, perceive, follow instructions, and implement complex instructions in a realistic environment (as opposed to virtual micro-worlds like that " inhabited " by SHRDLU [see Section 2.1]). SHAKEY has now retired from active service and lives in the Robot Hall of Fame at Carnegie Mellon University in Pittsburgh, Pennsylvania.

Figure 14.1 depicts one of the physical environments in which SHAKEY operated. The name of each room begins with an " R. " " RMYS " is a mystery room — that is, SHAKEY has no information about its contents. Doorway names begin with a " D " and are labeled in a way that makes clear which rooms they are connecting. " DUNIMYS, " for example, labels the door between RUNI (where SHAKEY starts) and RMYS. The environment is empty, except for three boxes located in RCLK (the room with the clock).

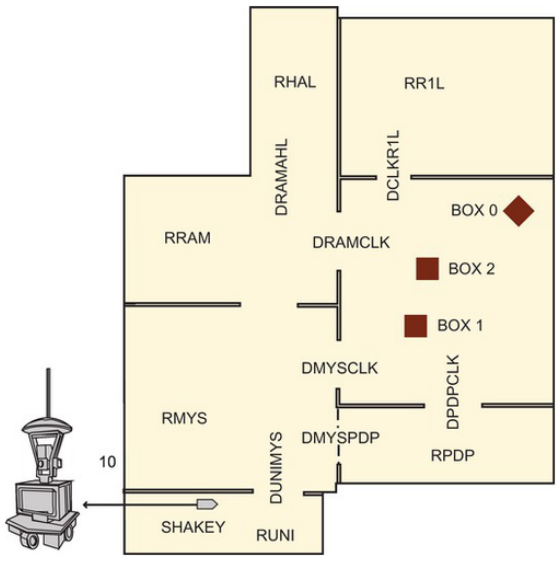


Figure 14.1 A map of SHAKEY ′ s physical environment. Each room has a name. The room containing the boxes is called *RCLK* (an abbreviation for *Room with Clock*). The total environment measures about 60 ft by 40 ft.

(Adapted from Nilsson 1984 .)

In thinking about SHAKEY, the first place to start is with the physical structure itself, which is shown in Figure 14.2 . The software that allows SHAKEY to operate is not actually run on the robot itself. It was run on a completely separate computer system that communicated by radio with SHAKEY (the radio antenna can be seen in the photo, extending from the top of the structure).

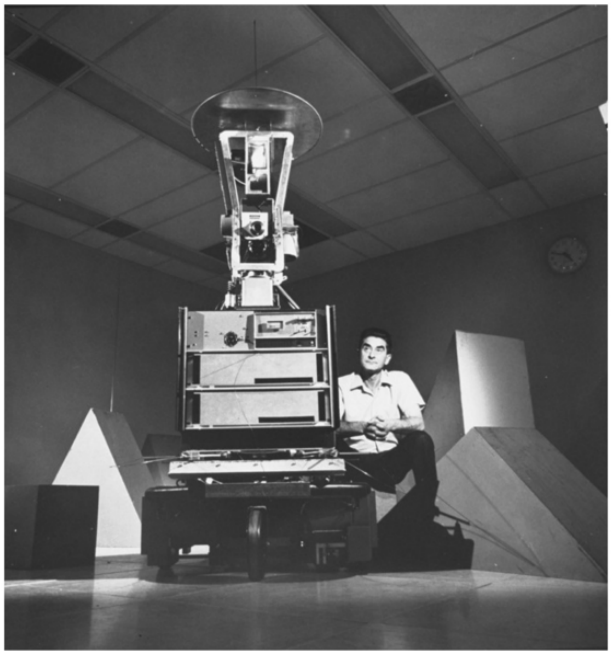


Figure 14.2 A photograph of SHAKEY the robot.

(Photograph by Ralph Crane/The LIFE Picture Collection/Getty Images.)

The programs that run SHAKEY are examples of *logic programming.* They incorporate a basic model of the environment, together with a set of procedures for updating the model and for acting on the environment.

SHAKEY ′ s basic model is given by a set of statements in the *first-order predicate calculus.* (This is the logical language that allows us to talk about particular objects having particular properties and also permits us to formulate generalizations either about all objects or about at least one object.) These statements are in a basic vocabulary that contains names for the objects in the robot ′ s world — doors, blocks, walls, and so on — as well as predicates that characterize the properties those objects can have. The vocabulary also contains a name for SHAKEY and predicates that describe the robot ′ s state — where it is, the angle at which its head is tilted,

and so on. The software that SHAKEY uses to plan and execute its actions exploits this same vocabulary, supplemented by terms for particular actions.

SHAKEY ′ s Software I: Low-Level Activities and Intermediate-Level Actions

To understand SHAKEY ′ s software, we need to go back to Chapter 1 , where we looked at Lashley ′ s influential and innovative ideas about actions. Reacting against the behaviorist idea that actions could be viewed as linked chains of responses, Lashley argued that many complex behaviors result from prior planning and organization. These behaviors are organized hierarchically (rather than linearly). An overall plan (say, walking over to the table to pick up the glass) is implemented by simpler plans (the walking plan and the reaching plan), each of which can be broken down into simpler plans, and so on. Ultimately, we arrive at basic actions that don ′ t require any planning. These basic actions are the components from which complex behaviors are built.

SHAKEY ′ s software packages are built around this basic idea, as illustrated in Table 14.1 , which shows how we can think about SHAKEY as a system with five different levels. The bottom level is the hardware level, and there are four different levels of software. The software levels are hierarchically organized. Each level of software controls a different type of behavior. Going up the hierarchy of software takes us up the hierarchy of behavior.

Table 14.1 SHAKEY ′ S five levels

LEVEL	FUNCTION	EXAMPLES
1. Robot vehicle and connections to user programs	To navigate and interact physically with a realistic environment	See the photograph of SHAKEY in Figure 14.2
2. Low-level actions (LLAs)	To give the basic physical capabilities of the robot	ROLL (which tells the robot to move forward by a specified number of feet) and TILT (which tells the robot to tilt its head upward by a specified number of degrees)
3. Intermediate-level actions (ILAs)	Packages of LLAs	PUSH (OBJECT, GOAL, TOL), which instructs the robot to push a particular object to a specified goal, with a specified degree of tolerance
4. STRIPS	A planning mechanism constructing MACROPS (sequences of ILAs) to carry out specific tasks	A typical MACROP might be to fetch a block from an adjacent room
5. PLANEX	Executive program that calls up and monitors individual MACROPS	PLANEX might use the sensors built into the robot to determine that the block can only be fetched if SHAKEY pushes another block out of the way first—and then invoke a MACROP to fetch a block

The low-level actions (LLAs) are SHAKEY ′ s basic behaviors — the building blocks from which everything that it does is constructed. The LLAs exploit the robot ′ s basic physical capabilities. So, for example, SHAKEY can move around its environment by rolling forward or backward. It can take photos with the onboard camera, and it can move its head in two planes — tilting it up and down and panning it from side to side. There are LLAs corresponding to all of these abilities.

In themselves, however, LLAs are not much use for problem solving and acting. SHAKEY ′ s designers needed to build a bridge between high-level commands (such as the command to fetch a block from a particular room) and the basic movements that SHAKEY can use to carry out that command. The first level of organization above LLAs comes with intermediate-level actions (ILAs). The ILAs are essentially action routines — linked sequences of LLAs that SHAKEY can call on in order to execute specific jobs, such as navigating to another room or turning toward a goal. Table 14.2 shows some ILAs.

Table 14.2 SHAKEY ′ S intermediate-level routines

ILA	ROUTINES CALLED	COMMENTS
PUSH3	PLANOBMOVE*, PUSH2	Can plan and execute a series of PUSH2s
PUSH2	PICLOC*, OBLOC*, NAVTO, ROLLBUMP, PUSH1	Check if object being pushed slips off
PUSH1	ROLL*	Basic push routine; assumes clear path
GETTO	GOTOROOM, NAVTO	Highest-level go-to routine
GOTOROOM	PLANTOUR*, GOTOADJROOM	Can plan and execute a series of GOTOADJROOMs
GOTOADJROOM	DOORPIC*, ALIGN, NAVTO, BUMBLETHRU	Tailored for going through doorways
NAVTO	PLANJOURNEY*, GOTO1	Can plan and execute a trip within one room
GOTO1	CLEARPATH*, PICDETECTOB*, GOTO	Recovers from errors due to unknown objects
GOTO	PICLOC*, POINT, ROLL2	Executes single straight-line trip
POINT	PICTHETA*, TURN2	Orients robot toward goal
TURN2	TURNBACK*, TURN1	Responds to unexpected bumps
TURN1	TURN*	Basic turn routine; expects no bumps
ROLL2	ROLLBACK*, ROLL1	Responds to unexpected bumps
ROLL1	ROLL*	Basic roll routine that expects no bumps
ROLLBUMP	ROLLBACK*, ROLL1	Basic roll routine that expects a terminal bump

ILAs are not just chains of LLAs (against the behaviorist idea that complex actions are chained sequences of basic responses). They can recruit other ILAs. So, for example, the GETTO action routine takes SHAKEY to a specific room. This action routine calls on the NAVTO routine for navigating around in the room SHAKEY is currently in, as well as the GOTOROOM routine, which takes SHAKEY to the room it is aiming for. Of course, SHAKEY can only move from any room to an adjacent room, so the GOTOROOM routine is built up from the GOTOADJROOM routine.

Nothing that we have seen so far would properly be described as problem solving — or indeed, to go back to Newell and Simon ′s physical symbol system hypothesis, as intelligent action.

SHAKEY ′s Software II: Logic Programming in STRIPS and PLANEX

The real innovation in SHAKEY ′s programming came with the STRIPS planner (*STRIPS* is an acronym for *Stanford Research Institute Problem Solver*). The STRIPS planner (which, as it happens, was fairly closely related to Newell and Simon ′s General Problem Solver [GPS]) allows SHAKEY to do things that look much more like reasoning about its

environment and its own possibilities for action. What STRIPS does is translate a particular goal statement into a sequence of ILAs.

SHAKEY has an axiomatic model of its environment. The axioms are well-formed formulas in the predicate calculus, built up from a basic vocabulary for describing SHAKEY and its environment. These formulas describe both SHAKEY ' s physical environment and its own state. The model is updated as SHAKEY moves around and acts on the environment.

The tasks that SHAKEY is given are also presented as formulas in the predicate calculus. The predicate calculus is a tool for deduction, and what SHAKEY does, in essence, is to come up with a deduction that has its goal as its conclusion.

We can think about SHAKEY ' s planning process as involving a tree search. (Think back to the decision trees that we looked at in Section 12.1 .) The first node (the top of the tree) is SHAKEY ' s model of the current environment. Each branch of the tree is a sequence of ILAs. Each node of the tree is an updated model of the environment.

If the goal formula can be deduced from the updated model at a given node, then STRIPS has solved the problem. What it then does is instruct SHAKEY to follow the sequence of ILAs described in the branch that leads to the solution node. SHAKEY does this, updating its model of the environment as it goes along.

There is no guarantee that this will always get SHAKEY to where it wants to go. The goal might not be attainable. Its model of the environment might not be correct. Someone might have moved the block without telling SHAKEY (and in fact, researchers at the SRI did precisely that to see how SHAKEY would update its model). This is where the PLANEX level comes into play. The PLANEX software monitors the execution of the plan. So, for

example, it contains an algorithm for calculating the likely degree of error at a certain stage in implementing the task (assuming that executing each ILA would introduce a degree of " noise " into SHAKEY ' s model of the environment). When the likely degree of error reaches a certain threshold, PLANEX instructs SHAKEY to take a photograph to check on its position. If a significant error is discovered, then PLANEX makes corresponding adjustments to the plan.

SHAKEY is a great example of the physical symbol system hypothesis in action. The physical symbol structures are well-formed formulas in the predicate calculus. These symbols give SHAKEY ' s model of the environment, as well as the goals and subgoals that SHAKEY is trying to achieve. And the physical symbol structures are manipulated and transformed by the STRIPS and PLANEX algorithms.

Moreover, SHAKEY clearly illustrates the heuristic search hypothesis (see Chapter 4). The hypothesis says that intelligent problem solving takes place by transforming and manipulating symbol structures until a solution structure is reached. The starting point is given by SHAKEY ' s model of the environment, together with the target formula that represents the desired end state. The permissible transformations are given by the STRIPS and PLANEX algorithms. And the problem is solved when the initial symbol structure has been transformed into a symbol structure from which the target formula can be deduced .

14.2 Situated Cognition and Biorobotics

The situated-cognition movement is similar in many ways to the dynamical systems approach, which we discussed in Chapter 6 . Cognitive scientists influenced by dynamical systems theory propose a new way of analyzing and predicting cognitive systems as coupled systems, focusing on variables evolving through state space in real time rather than representations. Instead of abstracting away from the physical details of how cognitive systems actually work, they suggest that those physical details can play all sorts of unsuspected but vitally important roles in determining how a cognitive system changes and evolves over time.

Situated-cognition theorists propose a broadly similar approach to robotics, moving away from the disembodied approach that we find in SHAKEY (with a sharp distinction between hardware and software) and toward design principles that exploit and leverage the robot ' s physical properties, as well as its embeddedness in real-life environment and its need to operate in real time.

The Challenge of Building a Situated Agent

Situated-cognition theorists argue that traditional cognitive science has never really come to terms with real-life problems and challenges in understanding cognition. This problem is particularly acute for the GOFAI approach to building artificial agents.

SHAKEY, which we have just looked at, is a classic example of GOFAI robotics. So, too, is Terry Winograd ' s SHRDLU program for natural language understanding, which we looked at in Section 2.1 . SHRDLU is a virtual robot, reporting on and interacting with a virtual micro-world. A good

entry point for the worries that situated-cognition theorists have about GOFAI is via a criticism often leveled at SHRDLU and other micro-world programs.

Here is a witty expression of the worry from philosopher and cognitive scientist John Haugeland (although he is not himself a promoter of the situated-cognition movement):

> SHRDLU performs so glibly only because his domain has been stripped of anything that could ever require genuine wit or understanding. Neglecting the tangled intricacies of everyday life while pursuing a theory of common sense is not like ignoring friction while pursuing the laws of motion; it ′ s like throwing the baby out with the bathwater. A round frictionless wheel is a good approximation of a real wheel because the deviations are comparatively small and theoretically localized: the blocks-world " approximates " a playroom more as a paper plane approximates a duck.
>
> (Haugeland 1985 : 190)

In other words, SHRDLU only works because its artificial micro-world environment has been stripped of all complexity and challenge.

SHRDLU does not have to work out what a block is, for example — or how to recognize one. There is very little " physical " challenge involved in SHRDLU ′ s (virtual) interactions with its micro-world environment because SHRDLU has built-in programs for picking up blocks and moving them around, and the robot hand is expressly designed for implementing those programs. Likewise, SHRDLU ′ s language-understanding achievements are partly a function of its artificially limited language and the highly circumscribed conversational context. The major problems in language

understanding (such as decoding ambiguity and working out what a speaker is really trying to say) are all factored out of the equation. Finally, SHRDLU is not autonomous — it is a purely reactive system, with everything it does being a response to explicit instructions.

In sum, SHRDLU is not properly *situated* in its environment, which is so constrained and devoid of meaning that it is positively misleading to take it as a starting point in thinking about human cognition.

The call for situated cognition, then, is a call for artificial intelligence (AI) to work on systems that have all the things that SHRDLU lacks — systems that are properly embodied and have real autonomy. These systems need to be embedded in something much more like the real world, with ambiguous, unpredictable, and highly complex social and physical contexts.

For situated-cognition theorists, SHAKEY is no more a situated agent than SHRDLU, even though it propels itself around a physical environment. They argue that the real work has already been done in writing SHAKEY ' s program. SHAKEY ' s world is already defined for it in terms of a small number of basic concepts (such as BOX, DOOR, and so forth). Its motor repertoire is built up out of a small number of primitive movements (such as ROLL, TILT, and PAN). The problems that SHAKEY is asked to solve are presented in terms of these basic concepts and primitive movements (such as when SHAKEY is asked to fetch a BOX).

The robot has to work out a sequence of basic movements that will fulfill the command, but that is not the same as a real agent solving a problem in the real world. SHAKEY already has the basic building blocks for the solution. But working out what the building blocks are is perhaps the most difficult part of real-world problem solving. Like SHRDLU, SHAKEY can only operate successfully in a highly constrained environment. Situated-

cognition theorists are interested in building agents that will be able to operate successfully even when all those constraints are lifted .

Situated Cognition and Knowledge Representation

Rodney Brooks is a very influential situated-cognition theorist, and his paper " Intelligence without Representation " is something of a manifesto for situated-cognition theorists. Brooks has an interesting diagnosis of where traditional AI has gone wrong.

Brooks points out that classical AI depends crucially on trimming down the type and number of details that a cognitive system has to represent. Here is his example in his own words:

> Consider chairs, for example. While these two characterizations are true
>
> (CAN (SIT-ON PERSON CHAIR))
>
> and
>
> (CAN (STAND-ON PERSON CHAIR))
>
> there is really much more to the concept of a chair. Chairs have some flat (maybe) sitting place, with perhaps a back support. They have a range of possible sizes, and a range of possibilities in shape. They often have some sort of covering material — unless they are made of wood, metal or plastic. They sometimes are soft in particular places. They can come from a range of possible styles. In sum, the concept of what a chair is hard to characterize simply. There is certainly no AI vision program that can find arbitrary

> chairs in arbitrary images; they can at best find one particular type of chair in arbitrarily selected images.
>
> (Brooks 1991 : 399)

Recognizing and interacting with chairs is a complicated business. But the programmer can remove the complications more or less with a stroke — simply by programming into the system a very narrow characterization of what a chair is. The beauty of doing this is that it can make certain types of chair interactions very simple.

If, to continue with Brooks ' s example, the system has to solve a problem with a hungry person seated on a chair in a room with a banana just out of reach, then the characterization in the program is just what ' s required. But of course, if the system solves the problem, then this is largely because it has been given all and only the right sort of information about chairs — and because the problem has been presented in a way that points directly to a solution! Here is Brooks again:

> Such problems are never posed to AI systems by showing them a photo of the scene. A person (even a young person) can make the right interpretation of the photo and suggest a plan of action. For AI planning systems, however, the experimenter is required to abstract away most of the details to form a simple description of atomic concepts such as PERSON, CHAIR, and BANANA.
>
> But this abstraction process is the essence of intelligence and the hard part of the problem being solved. Under the current scheme, the abstraction is done by the researchers, leaving little for the AI programs to do but search. A truly intelligent program would study the photograph, perform the abstraction itself, and solve the problem.

(Brooks 1991 : 399)

This gives us a much clearer view of what situated cognition is supposed to be all about. It ′ s not just a question of designing robots that interact with their environments. There are plenty of ways of doing this that don ′ t count as situated cognition. The basic idea is to develop AI systems and build robots that don ′ t have the solutions to problems built into them — AI systems and robots that can learn to perform the basic sensory and motor processes that are a necessary precondition for intelligent problem solving.

Biorobotics: Insects and Morphological Computation

Situated-cognition theorists, like dynamical systems theorists, believe that it pays to start small. Dynamical systems theorists often focus on relatively simple motor and cognitive behaviors, such as infant stepping and the A-not-B error. Cognitive scientists in situated robotics are often inspired by cognitively unsophisticated organisms. Insects are very popular. We can get the flavor from the title of another one of Rodney Brooks ′ s influential articles —“ Today the Earwig, Tomorrow Man? ”

Instead of trying to model highly simplified and scaled-down versions of “ high-level ” cognitive and motor abilities, situated cognition theorists think that we need to focus on much more basic and *ecologically valid* problems. The key is simplicity without simplification.

Insects solve very complex problems. Studying how they do this and then building models that exploit the same basic design principles will, according to theorists such as Brooks, pay dividends when it comes to understanding how human beings interact with their environment. We need to look at humans as scaled-up insects, not as scaled-down supercomputers.

Situated-cognition theorists stress the basic design principle of having direct links between perception and action — as opposed to the classical picture of distinct and semiautonomous subsystems that can be analyzed and modeled independently of each other. In a view like Marr ' s, for example, the visual system is an autonomous input - output system. It processes information completely independently of what will happen to that information further downstream. When we look at insects, however, we see that they achieve high degrees of " natural intelligence " through clever engineering solutions that exploit direct connections between their sensory receptors and their effector limbs.

Some researchers in this field have described what they are doing as *biorobotics.* The basic idea is usefully summarized in Figure 14.3 . Biorobotics is the enterprise of designing and building models of biological organisms that reflect the basic design principles built into those organisms.

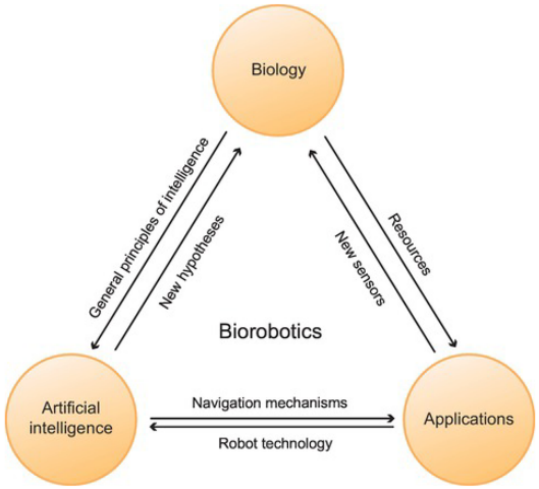


Figure 14.3 The organizing principles of biorobotics — a highly interdisciplinary enterprise.

(Courtesy of Dimitrios Lambrinos.)

Bioroboticists look to biology for insights into how insects and other simple organisms solve adaptive problems, typically those having to do with locomotion and foraging. They start with theoretical models. They then modify those models after seeing what happens when they are physically implemented in robots — robots whose construction is itself biologically inspired.

A famous example of biorobotics in action is the work of Edinburgh University ' s Barbara Webb on how female crickets locate males on the basis of their songs — what biologists call *cricket phonotaxis* .

Female crickets are extremely good at recognizing and locating mates from the song that they make. This might seem like a problem that can only be solved with very complex information processing — identifying the sound, working out where it comes from, and then forming motor commands that will take the cricket to the right place. Webb observed, however, that the physiology of the cricket actually provides a very clever solution, exploiting direct links between perception and action.

A remarkable fact about crickets is that they have their ears on their legs. As shown in Figure 14.4 , the cricket ' s ears are connected by a tube (the *tracheal tube*). This means that a single sound can reach each ear via different routes — a direct route (through the ear itself) and various indirect routes (via the other ear, as well as through openings in the tracheal tube known as *spiracles*). Obviously, a sound that takes the indirect route will take longer to arrive because it has farther to travel — and can ' t go faster than the speed of sound.

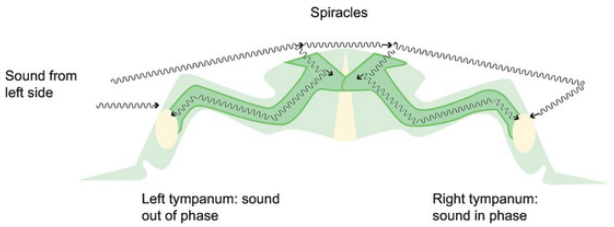


Figure 14.4 The cricket ′ s ears are on its front legs. They are connected to each other via a tracheal tube. The spiracles are small openings that allow air into the tracheal tube. The arrows show the different routes that a single sound can take to each ear.

(Adapted from Clark 2001 .)

According to Barbara Webb, cricket phonotaxis works because of two very basic design features of cricket anatomy. First, vibration is highest at the ear nearest the source of the sound (because the sound has traveled less far). Second, this vibration directly controls the cricket ′ s movements. Crickets are *hardwired* to move in the direction of the ear with the highest vibration (provided that the vibration is suitably cricket-like). There is no " direction-calculating mechanism, " no " male cricket identification mechanism, " and no " motor controller. "

Webb and her coworkers have used this model to build robot crickets that can actually perform a version of phonotaxis. In fact, not only can they find the sources of artificial cricket sounds, but they can even find real crickets by their sound.

One of the key design features of Webb ′ s robot cricket (reflecting how real crickets have evolved) is that the cricket ′ s body is a contributing factor in the computation. Cricket phonotaxis works by comparing two different signals. The availability of these two different signals is a direct function of the cricket ′ s bodily layout, as illustrated in Figure 14.4 . This can be seen as

an early example of what subsequently emerged as the *morphological computation* movement in robotics.

Morphology (in this context) is body shape. The basic idea behind morphological computation is that organisms can exploit features of body shape to simplify what might otherwise be highly complex information-processing tasks. Applying this idea to robotics means building as much of the computation as possible directly into the physical structure of the robot. Morphological computation is a research program for designing robots in which as much computation as possible is done for free.

Here are two other examples of morphological computation from the AI Lab in the Department of Informatics at the University of Zurich.

The first example is a fish called WANDA, illustrated in Figure 14.5 . WANDA is designed with only one degree of freedom. The only thing WANDA can do is wiggle its tail from side to side at varying amplitudes and frequencies — that is, WANDA can vary the speed and the degree with which its tail moves. And yet, because of the power of morphological computation, variation in tail wiggling allows WANDA to carry out the full range of fish movements in all three planes — up – down and left – right as well as forward.

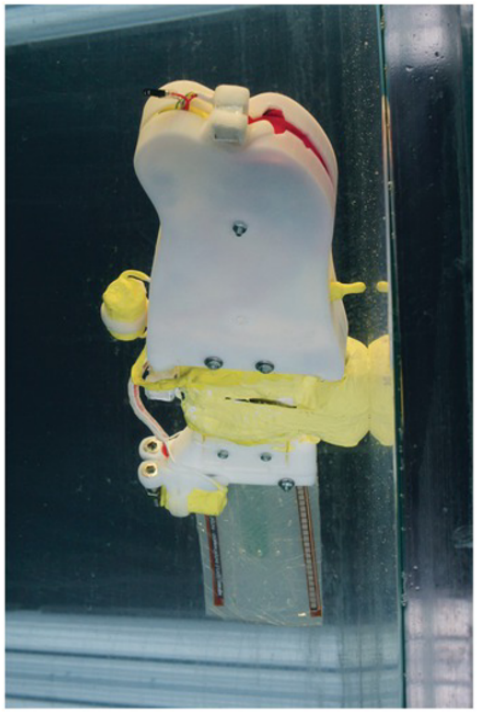


Figure 14.5 A robot fish called WANDA. All that WANDA can do is wiggle its tail fin. Yet, in an illustration of morphological computation, WANDA is able to swim upward, downward, and from side to side.

(Courtesy of Dimitrios Lambrinos.)

Part of the trick here is WANDA ′ s buoyancy, which is set so that slow tail wiggling will make it sink, whereas fast tail wiggling will make it rise. The other key design feature is the possibility of adjusting the zero point of the wiggle movement, which allows for movement to the left or right. Figure 14.6 shows WANDA swimming upward.

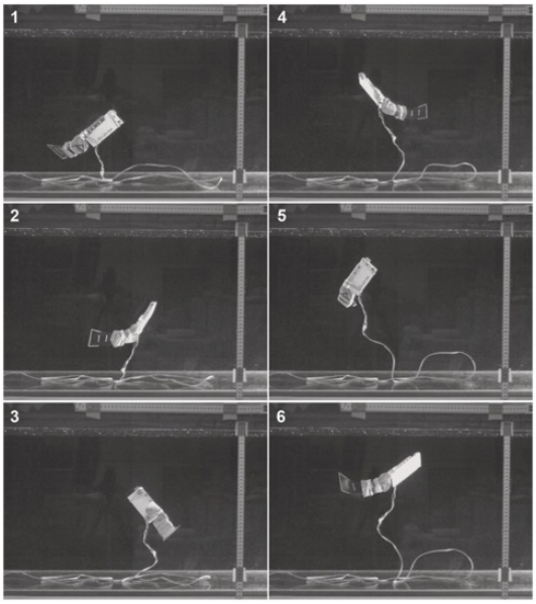


Figure 14.6 WANDA swimming upward.

(From Pfeifer et al. 2006 .)

A second example of morphological computation also comes from the realm of motor control. (We can think of both examples as ways of counterbalancing the appeal of the computational approach to motor control, which, as we saw in Chapter 6 , is also a target for dynamical systems theorists.) The robot hand devised by Hiroshi Yokoi in Figure 14.7 is designed to avoid the need for making explicit computations in carrying out grasping movements.

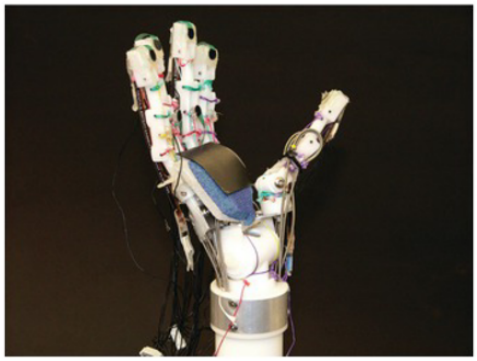


Figure 14.7 Another example of morphological computation: the robot hand designed by Hiroshi Yokoi. The hand is partly built from elastic, flexible, and deformable materials. The tendons are elastic, and both the fingertips and the spaces between the fingers are deformable. This allows the hand to adapt its grasp to the object being grasped.

(From Pfiefer et al. 2006.)

In the computational approach, grasping an object requires computing an object ′ s shape and configuring the hand to conform to that shape. Configuring the hand, in turn, requires sending a set of detailed instructions to the tendons and muscles determining the position of the fingers and palm. None of this is necessary, however, in controlling the Yokoi hand.

The hand is constructed from elastic and deformable materials (elastic tendons and deformable fingertips and spaces between the fingers). This morphology does the work that would otherwise be done by complex calculations within some sort of motor control unit. The hand ′ s flexible and elastic morphology allows it to adapt itself to the shape of the objects being grasped, as illustrated in Figure 14.8 .

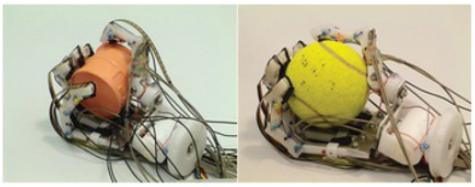


Figure 14.8 The Yokoi hand grasping two very different objects. In each case, the control is the same, but the morphology of the hand allows it to adapt to the shapes it encounters.

(From Bisig et al. 2008.)

As with the robot cricket example, most work in morphological computation has focused on the realm of motor control and sensorimotor integration. These are areas where traditional AI, and indeed, traditional cognitive science, have often been thought to be deficient. These are not cognitive tasks in any high-level sense. But they are often thought to require information processing, which is why they come into the sphere of cognitive science.

The real question, though, must be how the type of insights that we can find in biorobotics and morphological computation can be integrated into models of more complex agents. Some very suggestive ideas come from the field of behavior-based robotics, to which we turn in the next section .

14.3 From Subsumption Architectures to Behavior-Based Robotics

Rodney Brooks has provided a general AI framework for thinking about some of the agents discussed in the previous section. Webb ' s robot crickets are examples of what Brooks calls *subsumption architectures.*

Subsumption architectures are organized very differently from the modular architectures that we ′ ve been focusing on so far (see Chapter 8 , for example) and that are exemplified in SHAKEY. Subsumption architectures are made up of layers, and the layers are built up from behaviors.

The bottom level of the architecture is composed of very simple behaviors. Brooks ′ s favorite example is obstacle avoidance, which is obviously very important for mobile robots (and living organisms). The obstacle-avoidance layer directly connects perception (sensing an obstacle) to action (either swerving to avoid the obstacle or halting where the obstacle is too big to go around).

Whatever other layers are built into the subsumption architecture, the obstacle-avoidance layer is always online and functioning. This illustrates another basic principle of subsumption architectures. The layers are autonomous and work in parallel. There may be a " higher " layer that, for example, directs the robot toward a food source. But the obstacle-avoidance layer will still come into play whenever the robot finds itself on a collision course with an obstacle. This explains the name — the higher layers *subsume* the lower layers without replacing or overriding them.

This makes it easier to design creatures with subsumption architectures. The different layers can be grafted on one by one. Each layer can be exhaustively debugged before another layer is added. And the fact that the layers are autonomous means that there is much less chance that adding a higher layer will introduce unsuspected problems into the lower layers. This is obviously an attractive model for roboticists. It is also explicitly based on thinking about how evolution works.

Subsumption Architectures: The Example of Allen

Rodney Brooks ' s lab at the Massachusetts Institute of Technology (MIT) has produced many robots with subsumption architectures exemplifying these general principles. One of the first was Allen, illustrated in Figure 14.9 .

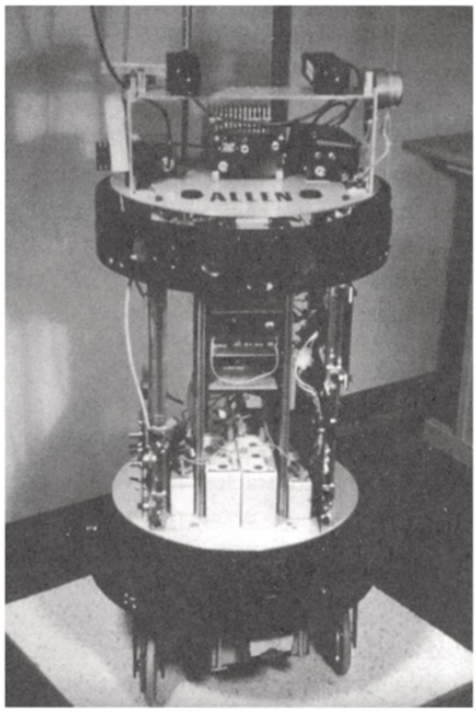


Figure 14.9 Rodney Brooks ' s robot Allen, his first subsumption-architecture robot.

(From Haugeland 1997 .)

At the hardware level, Allen looks like any GOFAI robot, not very dissimilar to SHAKEY, for example. At the software level, though, Allen is a subsumption architecture, built up in the standard layered manner. Over time, more and more layers were added to Allen ' s basic architecture. The first three layers are depicted in Figure 14.10 .

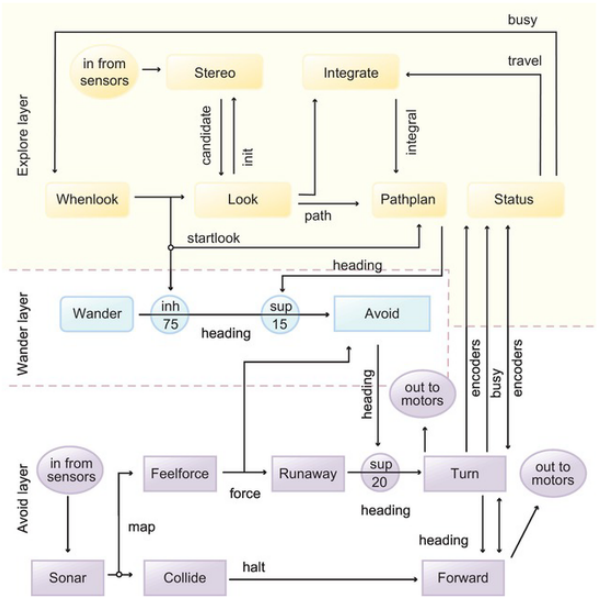


Figure 14.10 The layers of Allen ′s subsumption architecture. Allen has a three-layer architecture. The layers communicate through mechanisms of inhibition (inh) and suppression (sup).

(From Haugeland 1997 .)

The most basic layer is the obstacle-avoidance layer, which, as the diagram shows, is itself built up from distinct subsystems. The COLLIDE subsystem scans the sensory input for obstacles. It sends out a halt signal if it detects one. The FEELFORCE system works out the overall force acting on the robot (using information from the sensors and the assumption that objects function as repulsive forces). These feed into systems responsible for steering the robot — systems that are directly connected to the motor effectors.

In the middle layer, the WANDER component generates random paths for Allen ′s motor system, whereas the AVOID component feeds back down into the obstacle-avoidance layer to ensure that following the random path does not lead Allen to crash into anything. Allen is actually pretty successful

at this. The robot can successfully navigate environments with both stationary obstacles and other moving objects.

But Allen is not just wandering around at random. The subsystems in the top (*explorer*) layer allow Allen to pursue goals in a self-directed way. Using input from the sensory systems, they allow Allen to plan routes toward specific visible locations. As the wiring diagram in Figure 14.10 shows, the PATHPLAN subsystem feeds into the AVOID subsystem. This allows for the plan to be modified as the robot is actually moving toward the goal.

Drawing all this together, subsumption architectures have three basic features:

- *Incremental design:* Subsumption architecture robots are built to mimic how evolution seems to work. New subsystems are grafted on in layers without changing the existing subsystems.
- *Semiautonomous subsystems* : The subsystems operate relatively independently of each other, although some subsystems are set up to override others. The connections between the subsystems are hardwired. There is typically no central " controller, " comparable to STRIPS and PLANEX in SHAKEY
- *Direct perception – action links:* Subsumption architectures trade, as much as possible, on subsystems that deliver immediate motor responses to sensory input. They are designed for real-time control of action.

The contrast with traditional AI approaches is sharp. Traditional AI robots such as SHAKEY are designed in a very top-down way. There is typically a central planner maintaining a continuously updated model of the world,

updated by incorporating information received through its sensors. The planner uses this model of the world to work out detailed action plans, which are transmitted to the effectors. The action plans tend to be multistage and leave little scope for modification.

Proponents of GOFAI robotics are likely to say that the basic features of subsumption architectures are very good design principles for robots that are intended to be no more than mechanical insects — basically capable only of moving around the environment and reacting in simple ways to incoming stimuli. But subsumption architectures are not going to help us with complex, intelligent behavior.

Recall the phrasing of the physical symbol system hypothesis, which we looked at in detail in Chapter 4 and is the dominant theoretical framework for GOFAI robotics. The physical symbol system hypothesis is a hypothesis about the necessary and sufficient conditions of *intelligent action* . But how intelligent are Allen and the robot crickets and cockroaches that bioroboticists have developed?

There are different ways in which a situated-cognition theorist might try to respond to this challenge. One way is to try to combine the two approaches. There are hybrid architectures that have a subsumption architecture for low-level reactive control, in combination with a more traditional central planner for high-level decision-making. So, for example, Jonathan Connell, a researcher at IBM ′ s T. J. Watson Research Center in Yorktown Heights, New York, has developed a three-level hybrid architecture that he calls SSS.

It is easy to see where the acronym comes from when we look at what each of the layers does. SSS contains the following:

- A **s** ervo-based layer that controls the robot ′ s effectors and processes raw sensory data
- A **s** ubsumption layer that reacts to processed sensory input by configuring the servo-based layer (as is standard in a subsumption architecture, the different subsystems are organized in a strict precedence hierarchy)
- A **s** ymbolic layer that maintains complex maps of the environment and is capable of formulating plans; the symbolic layer configures the subsumption layer

The hybrid-architecture approach abandons some of the basic ideas behind situated cognition and biorobotics. To return to a phrase used earlier, situated-cognition theorists like to think of sophisticated cognitive systems as scaled-up insects, whereas GOFAI theorists think of them as scaled-down supercomputers. The hybrid-architecture approach, as its name suggests, looks for a middle way — it sets out to build scaled-up insects with scaled-down supercomputers grafted onto them.

But some situated-cognition theorists have tried to meet the challenge without compromising on the basic principles of situated cognition. *Behavior-based robotics* moves beyond basic subsumption architectures in a way that tries to build on the basic insights of the situated-cognition movement .

Behavior-Based Robotics: TOTO

Behavior-based architectures are designed to be capable of representing the environment and planning complex actions.

Subsumption architectures (and insects, for that matter) are purely reactive — they are designed to respond quickly to what is happening around them. These responses are typically fairly simple — such as changing the robot ' s direction or putting it into reverse when a collision is anticipated. These responses tend to be explicitly programmed into the system.

Behavior-based robots, in contrast, are capable of more complex behaviors that need not be explicitly specified within the system. These are sometimes called *emergent* behaviors (because they emerge from the operation and interaction of lower-level behaviors). This additional cognitive sophistication is gained without a symbol-using central planner.

Behavior-based architectures are typically built up from semiautonomous subsystems in a way that mimics the incremental approach that evolution seems to take. They have two key features, one that they share with subsumption architectures and one that sets them apart:

- *Real-time functioning:* Like subsumption architectures, behavior-based architectures are designed to operate in real time. They make plans on a timescale that interfaces directly with the robot ' s movements through the environment. This contrasts with symbolic planners and hybrid architectures, where planning is done offline and then needs to be integrated with the robot ' s ongoing behavior.
- *Distributed representations* : Behavior-based architectures represent their environments and use those representations in planning actions. This distinguishes them from most subsumption architectures. But unlike symbolic and hybrid architectures, those representations are not centralized or centrally manipulated. There is no central planning

system that gathers together all the information that the robot has at its disposal.

We can appreciate how these features work by looking at two examples from the work of Maja Matarić, one of the pioneers of behavior-based robotics. Before looking at how she applies the behavior-based approach to programming collections of robots (*multiagent programming*), we will explore how behavior-based robotics works for single robots.

Behavior-based architectures typically incorporate both *reactive rules* and *behaviors* . Subsumption architectures are basically built up from reactive rules. A reactive rule might, for example, tell the robot to go into reverse when its sensors detect a looming object. The reactive rules exploit direct perception – action links. They take inputs from the robot's sensors and immediately send instructions to the robot's effectors.

Behaviors are more complex. Matarić defines a behavior as *a control law that satisfies a set of constraints to achieve and maintain a particular goal.* The relevant constraints come both from the sensed environment (which might include other robots) and from the robot itself (e.g., its motor abilities).

So, the challenge for behavior-based robotics is to find a way of implementing behaviors in a mobile agent without incorporating a symbolic, central planner. Matarić's robot TOTO, which she designed and constructed together with Rodney Brooks, illustrates how this challenge can be met for a very specific navigation behavior. This is the behavior of finding the shortest route between two points in a given environment.

Matarić and Brooks were inspired by the abilities of insects such as bees to identify shortcuts between feeding sites. When bees travel from their hive, they are typically capable of flying directly to a known feeding site

without retracing their steps. In some sense, they (and many other insects, foraging animals, and migrating birds) are constructing and updating maps of their environment. This is a classic example of an apparently complex and sophisticated behavior being performed by creatures with very limited computational power at their disposal — exactly the sort of thing that behavior-based robotics is intended to model.

TOTO is designed to explore and map its environment (an office-like environment where the principal landmarks are walls and corridors) so that it can plan and execute short and efficient paths to previously visited landmarks. TOTO has a three-layer architecture. The first layer comprises a set of reactive rules. These reactive rules allow it to navigate effectively and without collisions in its environment. The second layer (the landmark-detector layer) allows TOTO to identify different landmarks. In the third layer, information about landmarks is used to construct a distributed map of the environment. This map is topological rather than metric. It simply contains information as to whether or not two landmarks are connected — but not as to how far apart they are. TOTO uses the topological map to work out, in real time, the shortest path back to a previously visited landmark (i.e., the path that goes by the smallest number of landmarks).

One of TOTO ′s key features is that its map is distributed (in line with the emphasis within behavior-based robotics on distributed representations), and the processing works in parallel. There is no single data structure representing the environment. Instead, each landmark is represented by a procedure that categorizes the landmark and fixes its compass direction.

The landmark procedures are all linked together to form a network. Each node in the network corresponds to a particular landmark, and if there is a direct path between two landmarks, then there is an edge connecting

them in the network. This network is TOTO ′ s topological map of the environment. It is distributed because it exists only in the form of connections between separate landmark procedures.

Behavior-based roboticists do not object to representations per se. They recognize that any robot capable of acting in complex ways in a complex environment must have some way of storing and processing information about its environment. Their real objection is to the idea that this information is stored centrally and processed symbolically.

TOTO constantly expands and updates its network as it moves through the environment detecting new landmarks. This updating is done by activation spreading through the network (not dissimilar to a connectionist network). When the robot is at a particular landmark, the node corresponding to that landmark is active. It inhibits the other nodes in the network (which is basically what allows TOTO to know where it is), and at the same time, it is spreading positive activation (expectation) to the next node in the direction of travel (which allows TOTO to work out where it is going).

This distributed map of the environment leaves out a lot of important information (about distances, for example). But for that very reason, it is flexible, robust, and most importantly, very quick to update.

Matari ć and Brooks designed an algorithm for TOTO to work out the shortest path between two nodes on the distributed map. The algorithm works by spreading activation. Basically, the active node (which is TOTO ′ s current location) sends a call signal to the node representing the target landmark. This call signal gets transmitted systematically through the network until it arrives at the target node. The algorithm is designed so that the route that the call signal takes through the network represents the shortest

path between the two landmarks. Then TOTO implements the landmark procedures lying on the route to navigate to the target landmark.

TOTO is a nice example of the key features of behavior-based robotics. TOTO is not simply a reactive agent, like Barbara Webb ′ s robot cricket. Nor does it have a central symbolic planner, like Jonathan Connell ′ s SSS. It has a distributed map of the environment that can be directly exploited to solve navigational problems. The basic activation-spreading mechanisms used for creating and updating the map are the very same mechanisms used for identifying the shortest paths between two landmarks. The mechanisms are somewhat rough-and-ready. But that is what allows them to be used efficiently in the real-time control of behavior — which, after all, is what situated cognition is all about .

Multiagent Programming: The Nerd Herd

Multiagent programming is highly demanding computationally, particularly if it incorporates some sort of centralized planner or controller. A central planner would need to keep track of all the individual robots, constantly updating the instructions to each one to reflect the movements of others — as well as the evolution of each robot ′ s own map of the environment. The number of degrees of freedom is huge.

The multiagent case presents, in a very stark way, a fundamental challenge of robotics: How to design a system that can reason about its environment without a complete combinatorial explosion? It is very instructive to see what happens when the challenge is tackled through the behavior-based approach.

Matari ć built a family of twenty mobile robots — the so-called Nerd Herd, illustrated in Figure 14.11 . Each robot was programmed with a set of basis behaviors (see Table 14.3). These basis behaviors are the building blocks for more complex *emergent* behaviors that were not explicitly programmed into the robots.

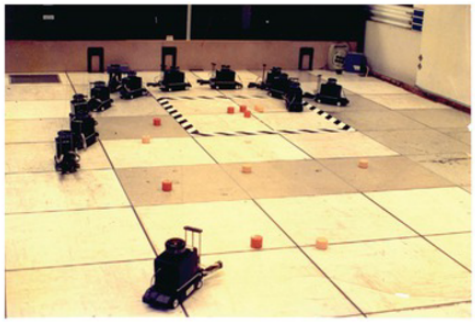


Figure 14.11 The Nerd Herd, together with the pucks that they can pick up with their grippers.

(Reprinted with permission of Maja Matari ć .)

Table 14.3 The five basis behaviors programmed into Matari ć' s Nerd Herd robots

Safe-wandering	Ability to move around while avoiding collisions with robots and other objects
Following	Ability to move behind another robot, tracing its path
Dispersion	Ability to maintain a *minimum* distance from other robots
Aggregation	Ability to maintain a *maximum* distance from other robots
Homing	Ability to find a particular region or location

The five basis behaviors can be combined in two ways. The first way is through summation. The outputs from two or more behaviors are summed together and channeled toward the relevant effector (e.g., the wheels of the robot). This works because all of the behaviors have the same type of output. They all generate velocity vectors, which can easily be manipulated mathematically. The second combination is through switching. Switching inhibits all behaviors except one.

Each basis behavior is programmed at the level of the individual robot. None is defined for more than one robot at a time, and there is no communication between robots. Matari ć found, however, that combining the basis behaviors at the level of the individual robots resulted in emergent behaviors at the level of the group.

So, for example, the Nerd Herd could be made to display flocking behavior by summing basis behaviors in each individual robot. The group flocked together as a whole if each robot ' s control architecture summed the basis behaviors of dispersion, aggregation, and safe-wandering. Adding in homing allowed the flock to move together toward a particular goal.

The principal activity of the robots in the Nerd Herd is collecting little pucks. Each robot has grippers that allow it to pick the pucks up. Matari ć used the control technique of switching between different basis behaviors in order to generate the complex behavior of foraging. If the robot doesn ' t have a puck, then all the basis behaviors are inhibited except safe-wandering. If safe-wandering brings it too close to other robots (and hence to potential competitors), then the dominant behavior switches to dispersion. If the robot has a puck, then the control system switches over to homing, and the robot returns to base.

You may be wondering just how intelligent these complex behaviors really are. It is true that flocking and foraging are not explicitly programmed into the system. They are emergent in the sense that they arise from the interaction of basis behaviors. But the mechanisms of this interaction are themselves programmed into the individual robots using the combinatorial operators for basis behaviors. They are certainly not emergent in the sense of being unpredictable. And one might think that at least one index of

intelligence in robots or computers more generally is being able to produce behaviors that cannot simply be predicted from the wiring diagram.

It is significant, therefore, that Matari ć′ s behavior-based robots are capable of learning some of these complex behaviors without having them explicitly programmed. She showed this with a group of four robots very similar to those in the Nerd Herd. The learning paradigm she used was reinforcement learning. What is reinforced are the connections between the states a robot is in and the actions it takes.

The complex behavior of foraging is really just a set of condition – behavior pairs — if the robot is in a certain condition (e.g., lacking a puck), then it yields total control to a single behavior (e.g., safe-wandering). So, learning to forage is, in essence, learning these condition – behavior pairs. This type of learning can be facilitated by giving the robot a reward when it behaves appropriately in a given condition, thus reinforcing the connection between condition and behavior.

Matari ć worked with two types of reinforcement — reinforcement at the completion of a successful behavior and feedback while the robot is actually executing the behavior. Despite the complexity of the environment and the ongoing multiagent interactions, Matari ć found that her four robots successfully learned group-foraging strategies in 95 percent of the trials.

Obviously, it is a long way from groups of robots foraging for hockey pucks in a closed environment to anything recognizable as a human social interaction. But behavior-based robotics does at least give us a concrete example of how some of the basic insights behind the situated-cognition movement can be carried forward. Perhaps it is time to change Rodney Brooks ′ s famous slogan: " Yesterday the earwig. Today the foraging robot. Tomorrow man? "

Summary

After opening with the classic example of SHAKEY in GOFAI robotics, this chapter focused on new developments in designing and building artificial agents. In their criticisms of traditional GOFAI, situated-cognition theorists have been inspired by very simple systems such as insects. We then considered how these theoretical ideas have been translated into particular robotic architectures, focusing on the subsumption architectures developed by Rodney Brooks and on Maja Matari ć′ s behavior-based robotics.

Checklists

The robot SHAKEY is an example of how a physical symbol system can interact with a real physical environment and reason about how to solve problems.

1. SHAKEY has a model of its environment given by a set of sentences in a first-order logical language. This model is updated as SHAKEY moves around.

2. SHAKEY ′ s software is hierarchically organized into four different levels. At the most basic level are primitive actions (low-level actions [LLAs]). These LLAs are organized into action routines (intermediate-level actions [ILAs]). SHAKEY solves problems by constructing a sequence of ILAs that will achieve a specific goal.

3. The STRIPS problem-solving software is an example of logic programming. It explores the logical consequences of SHAKEY ′ s

model of its current environment in order to work out which ILAs can be applied in that environment.

4. STRIPS then works out how the model would need to be updated if each ILA were executed in order to develop a tree of possible ILA sequences. If one of the branches of the tree leads to the desired goal state, then SHAKEY implements the sequence of ILAs on that branch.

Situated-cognition theorists react against some of the fundamental tenets of GOFAI cognitive science and robotics.

1. Good old-fashioned artificial intelligence (GOFAI) programs such as SHRDLU and SHAKEY can interact (virtually) with their environments. But situated-cognition theorists argue that the only intelligence in these agents is put there by the programmers.

2. The world of a GOFAI robot is already defined for it in terms of a small number of basic concepts. Likewise for its motor repertoire. This avoids the real problems of decoding the environment and reacting to the challenges it poses.

3. Situated-cognition theorists think that instead of focusing on simplified and scaled-down versions of " high-level " tasks, cognitive scientists should look at how simple organisms such as insects solve complex but ecologically valid problems.

4. Biorobotics is the branch of robotics that builds models of biological organisms reflecting the basic design principles that have emerged in evolution (e.g., Barbara Webb ' s work on cricket phonotaxis).

Subsumption architectures are a powerful tool developed by situated-cognition theorists such as Rodney Brooks.

1. Subsumption architectures are not made up of functional subsystems in the way that modular architectures are. Instead, they are built up from layers of semiautonomous subsystems that work in parallel.

2. Subsumption architectures mimic how evolution might work. New systems are grafted on in layers that typically don ′ t change the design of the existing systems.

3. Subsumption architectures trade as much as possible on direct perception – action links that allow the online control of action.

Subsumption architectures do not typically have decision-making systems built into them. Problems of action selection are solved by predefined precedence relations among subsystems. Situated-cognition theorists have to work out a more flexible solution to the action-selection problem.

1. One approach is to develop a hybrid architecture, combining a subsumption architecture for low-level reactive control with a more traditional symbolic central planner for high-level decision-making.

2. Behavior-based robotics takes another approach. Behavior-based architectures (such as that implemented in TOTO) represent their environments and use those representations to plan actions. But these representations are not centralized or centrally manipulated.

3. In addition to reactive rules such as those in subsumption architectures, behavior-based robots have basis behaviors programmed

into them. These basis behaviors are more complex and temporally extended than reactive rules. They can also be combined.

4. Behavior-based robots can exhibit emergent behaviors that have not been programmed into them (e.g., the flocking and foraging behaviors displayed by Matari ć′ s Nerd Herd). Behavior-based robots have also been shown to be capable of learning these emergent behaviors through reinforcement learning.

Further Reading

SHAKEY is very well documented in technical reports published by the SRI. These can be downloaded at www.ai.sri.com/shakey/ . Technical Report 323 is particularly helpful. Also see the *Encyclopedia of Cognitive Science* entry on STRIPS (Nadel 2005).

The philosopher Andy Clark is a very clear expositor of situated cognition and biorobotics — see particularly *Being There* (1997) and Chapter 6 of Clark 2001 , as well as *Supersizing the Mind* (2008) and the discussion of that book in *Philosophical Studies* (2011). For more on morphological computation, see Pfeifer et al. 2006 (which includes the two examples in the text) and, more recently, Bernth et al. 2018 and Langer and Ay 2021 .

Clancey 1997 is a general survey of situated cognition from the perspective of an Al specialist. Several of Rodney Brooks ′ s influential papers are reprinted in his book *Cambrian Intelligence* (1999), which also contains some more technical papers on specific architectures. Brooks 1991 is also reprinted in Haugeland 1997 . For early versions of some of the

criticisms of GOFAI made by situated-cognition theorists, see Dreyfus 1977 . For a very different way of thinking about situated cognition (in terms of situatedness within a social environment), see Hutchins 1995 . The *Cambridge Handbook of Situated Cognition* (Robbins and Aydede 2008) is a useful and comprehensive resource, with a strong emphasis on the philosophical underpinnings of the situated-cognition movement. See also Miller and Feil-Siefer 2016 for the relation between situatedness and morphology. On embodied cognition, see Shapiro 2007 , Chemero 2009 , Adams and Aizawa 2010 , Shapiro 2011 , Anderson et al. 2012, and Lawrence Shapiro ' s chapter in Margolis et al. 2012 .

Arkin 1998 is a comprehensive textbook on behavior-based robotics. For a more programming-oriented survey, see Jones and Roth 2003 . Winfield 2012 is a more recent introduction. Maja Matari ć has written many papers on behavior-based robotics (see online resources). Matari ć 1997 and 1998 are good places to start. Readers interested in building their own mobile robots will want to look at her book *The Robotics Primer* (2007). Recent work on behavior-based robotics and robot swarms includes Tzafestas 2018 , Garcia-Aunon et al. 2019 , and Hasbach and Bennewitz 2021 .

The robot cricket made its first appearance in Webb 1995 . Shigeo Hirose was a pioneer of bio-inspired robotics, starting out with snakes, as described in his book *Biologically Inspired Robots* (Hirose 1993). For a more contemporary overview of bio-inspired robotics, see Ijspeert 2014 and Ryczko et al. 2020 . For examples of specific projects, see Libby et al. 2012 (tailed-assisted pitch control in lizards and dinosaurs), Koh et al. 2015 (water-jumping robots), and Graule et al. 2016 (insects perching and taking off) . Krichmar 2018 is an autobiographical review of neurorobotics.

Chapter 15

The Cognitive Science of Consciousness

◈

Overview

15.1 Information Processing without Conscious Awareness: Some Basic Data

- Consciousness and Priming
- Nonconscious Processing in Blindsight and Unilateral Spatial Neglect

15.2 So What Is Consciousness For?

- What Is Missing in Blindsight and Spatial Neglect
- Milner and Goodale: Vision for Action and Vision for Perception
- What Is Missing in Masked Priming

15.3 Two Types of Consciousness and the Hard Problem

15.4 Theories of Consciousness

- Higher-Order Theories
- The Global Workspace Theory
- The Global Neuronal Workspace Theory

15.5 Conclusion

Overview

Cognitive scientists typically work backward from observable behaviors and capacities to the information-processing mechanisms that generate those behaviors and support those capacities. This *third-person* perspective is broadly continuous with what physicists, chemists, and biologists do. Cognitive scientists are working with and trying to explain publicly observable phenomena — reaction times, levels of blood oxygen, verbal reports, and so forth.

But there is another perspective not yet discussed. This is the first-person perspective. Human cognitive agents have sensations. They experience the distinctive smell of a rose, the distinctive sound of chalk on a blackboard, the distinctive feel of cotton against the skin. They react emotionally to events and to each other. They regret the past and have hopes and fears for the future. From the first-person perspective, we have a rich *conscious* life, full of feelings, emotions, sensations, and experiences. These are all vital parts of what makes us human. How can we make sense of them within the information-processing model of the mind?

This tension between the first- and third-person perspectives makes consciousness an almost bipolar topic in contemporary cognitive science. On the one hand, we have many exciting experiments and creative theories aiming to understand what consciousness is and how it contributes to cognition. On the other, there are powerful arguments that it is impossible to give an information-processing model of consciousness. This chapter looks at both sides of the debate.

Section 15.1 explores the differences between conscious and nonconscious information processing, as revealed in priming experiments

and the behavior of patients with brain damage. Section 15.2 draws on these findings to explore theories about the function of consciousness. Then, in Section 15.3 , we look at arguments that functional approaches to consciousness cannot help us understand what is truly mysterious about consciousness — at best, they can shed light on those questions that are sometimes called the " easy " problems of consciousness. Section 15.4 presents the other side of the coin by reviewing two of the best-established approaches to the functional role of consciousness — the *higher-order thought theory* and the *global workspace theory.*

15.1 Information Processing without Conscious Awareness: Some Basic Data

One way to explore the nature of consciousness is to look at the types of information processing and problem solving that can take place without consciousness and compare them with those that seem to require consciousness. That will help us understand the function of consciousness — what it actually contributes to cognitive and affective life. We look at two techniques for studying information processing without consciousness — priming experiments and double dissociations in cognitive neuropsychology.

Consciousness and Priming

In a typical priming experiment, as illustrated in Figure 15.1 , subjects are exposed very briefly to some stimulus — an image on a screen, perhaps, or a sound. The exposure is short enough that the subjects do not consciously

register the stimulus. The experiment tests whether the exposure to the stimulus affects performance on subsequent tasks.

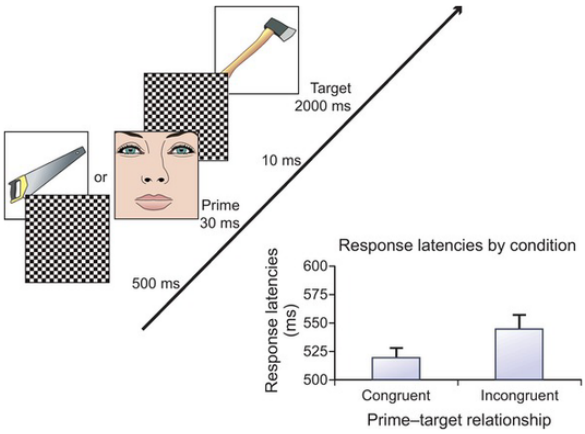


Figure 15.1 A typical priming experiment. The images above the arrow depict the sequence and timing of each stimulus when a tool is the target. The graph shows that people who were presented with a congruent prime were faster to identify the target than people who were presented with an incongruent prime.

(From Finkbeiner and Forster 2008 .)

In the experiment in Figure 15.1 , subjects are asked to categorize the target as either a face or a tool. There are two different types of primes. One type is congruent with the target (e.g., another tool if the target is a tool). The other is not congruent (e.g., a tool if the target is a face). The experiment measures the response latency (the time it takes the subject to classify the target correctly). As the graph illustrates, the experimenters found a significant priming effect for congruent prime – target pairs.

What does this reveal? Think about what counts as a congruent prime – target pair. Figure 15.1 gives one example — a saw and a hammer. These are congruent because they both fall under a single category. Noncongruent

prime – target pairs fall under different categories. So, what the priming effect appears to show is that the information processing required to carry out basic categorization can take place nonconsciously. The processing time for correctly classifying a congruent target is less than that for a noncongruent target, the standard explanation runs, because the subject is already thinking nonconsciously about the relevant category.

Some cognitive scientists have objected to the priming paradigm, worried about how one can show that primes really are invisible. A typical method of doing this is to identify a threshold by progressively lowering the presentation time of a stimulus until subjects identify it at chance rates. This is supposed to show that any stimulus presented for a length of time at or below the threshold will be nonvisible and nonconscious. But one problem with this is that the threshold of visibility can vary. There is some evidence that primes become more visible when they are followed by congruent targets. Varying the mask can also alter the threshold of visibility.

More recent studies have been very sensitive to these methodological issues, and the majority view now is that priming effects do occur, and hence that there are some kinds of nonconscious information processing. The crucial question is: How " smart " is this nonconscious information processing?

Priming effects have been identified in language processing, where primes and targets are both words. The most controversial experiments have focused on *semantic priming*. This is a priming effect that can only be explained through information processing about the meaning of words — as opposed, for example, to their phonology (how they are pronounced) or their orthography (how they are spelled).

There is interesting evidence for semantic priming from studies with bilingual subjects where the prime and target are in different languages, particularly where those languages are in very different scripts (Chinese and English, for example). Many studies have shown robust priming effects when subjects are asked to decide whether or not a target string of letters is a proper word or not (what is called the *lexical decision task*). Interestingly, the priming effect tends to occur only when the prime is in the dominant (first) language (L1) and the target is in the second language (L2).

This is significant because semantic processing is thought to be very high level. To go back to the distinction we looked at in Chapter 8 , semantic processing has typically been thought to be nonmodular (as opposed to processes such as phonological parsing, often thought to be modular). So, semantic priming is important because it seems to show that there can be information processing that is both nonmodular and nonconscious.

Exercise 15.1 Explain in your own words why the distinction between modular and nonmodular information processing is important for thinking about nonconscious information processing.

Nonconscious Processing in Blindsight and Unilateral Spatial Neglect

Cognitive neuropsychologists study cognitive disorders resulting from brain damage. The guiding idea is that we can work backward from what happens when things go wrong to how they function in the normal case. So, for

example, if in one type of brain damage we see ability A functioning more or less normally while ability B is severely impaired, then we can infer that in some sense, A and B are independent of each other — or, as cognitive neuropsychologists call it, we can infer a *dissociation* between them.

A *double dissociation* occurs when we have a dissociation in each direction — that is, in one disorder, we have ability A functioning normally with B significantly impaired, whereas in a second disorder, we have ability B functioning normally with A significantly impaired. Double dissociations provide stronger evidence that A and B are independent of each other.

> **Exercise 15.2** Explain in your own words why a double dissociation is a better sign of independence than a single dissociation.

Cognitive psychologists studying psychological disorders caused by brain trauma have identified surprisingly many tasks that can be carried out nonconsciously by patients with brain damage, even though they are typically performed with conscious awareness by normal subjects. In particular, we will look at two much-studied disorders — *unilateral spatial neglect* and *blindsight*.

Unilateral spatial neglect (also known as *hemiagnosia* or *hemineglect*) is relatively common and typically occurs after damage to the right hemisphere, particularly to the parietal and frontal lobes. Its defining feature is that patients lack awareness of sensory events on the contralesional side of space (on the opposite side of the world to the side of

the brain that is damaged). In the vast majority of cases, the neglected side is the left-hand side.

Italian neuropsychologists Eduardo Bisiach and Claudio Luzzatti in 1978 asked two patients with neglect to describe from memory the central square in Milan with the famous Duomo (cathedral). The patients were initially asked to describe the square as if they were standing in front of the Duomo. As predicted, the patients failed to describe the houses and shops on the left-hand side of the square (from their vantage point in front of the Duomo). Bisiach and Luzzatti then asked the patients to orient themselves differently so that they were imagining themselves on the edge of the square looking at the Duomo. Now the patients accurately described the houses and shops they had previously neglected, and they instead missed out the side of the square that they had previously described. Figure 15.2 shows further examples of typical visual deficits in patients with neglect.

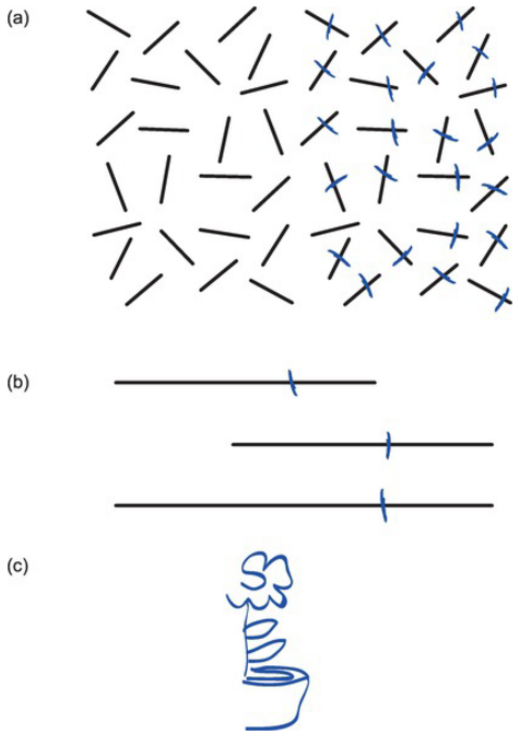


Figure 15.2 Examples of deficits found in patients with left spatial neglect (damage to the right hemisphere of the brain). (a) Patients with unilateral neglect typically fail to mark the lines on the contralesional (here, left) side of a sheet of paper. (b) Patients are asked to bisect each line. Their markings are typically skewed to the right, as if they do not see the leftmost segment. (c) Patients are either asked to draw something from memory or to copy another illustration placed in front of them. In both cases, patients with unilateral neglect tend to omit parts on the contralesional side.

(From Driver and Vuilleumier 2001 .)

Neglect also affects action. A patient with neglect might only shave or apply makeup to one side of their face, for example. Or they might eat only from one side of a plate.

Similarly, patients with blindsight typically report little to no awareness in one side of their visual field. They have what is called a *scotoma* (a portion of the visual field with very diminished visual acuity). Both patients with blindsight and those with neglect report themselves to be unaware of what is going on in part of their visual field. The etiology (cause) is different, however. The impairment in blindsight is typically due to lesions in the primary visual cortex (V1, or the striate cortex).

One challenge in exploring the residual abilities of patients with blindsight is that they will often find the experiments absurd. Ernst Pöppel, whose important 1973 article coauthored with Douglas Frost and Richard Held was one of the first to study blindsight, reported a patient irritatedly saying, " How can I look at something that I haven ' t seen? " when asked to direct his eyes to a target in his blind field.

To overcome this challenge, experimenters have used nonverbal *forced-choice tests* . Patients are forced to guess in situations where they feel that they have no basis for making a judgment or performing an action. The choices are usually binary — is the stimulus moving or stationary? Is it high or low in the visual field? Is it horizontal or vertical?

Exercise 15.3 Explain in your own words the rationale for using forced-choice tests with patients with blindsight.

Experimenters often find that patients with blindsight perform significantly better than chance, even when the patients describe themselves as guessing (and so would be expected to perform at chance levels). There is strong evidence that patients with blindsight can localize unseen stimuli

in the blind field, that they can discriminate orientation, and that they can detect moving and stationary figures randomly interspersed with blank trials.

Patients with blindsight are also capable of some types of form perception. Here is an example from a striking set of experiments performed by Ceri Trevethan, Arah Sahraie, and blindsight pioneer Larry Weiskrantz, working with a patient known by his initials, D.B. Figure 15.3 depicts line drawings of animals that were presented within D.B.'s blind field.

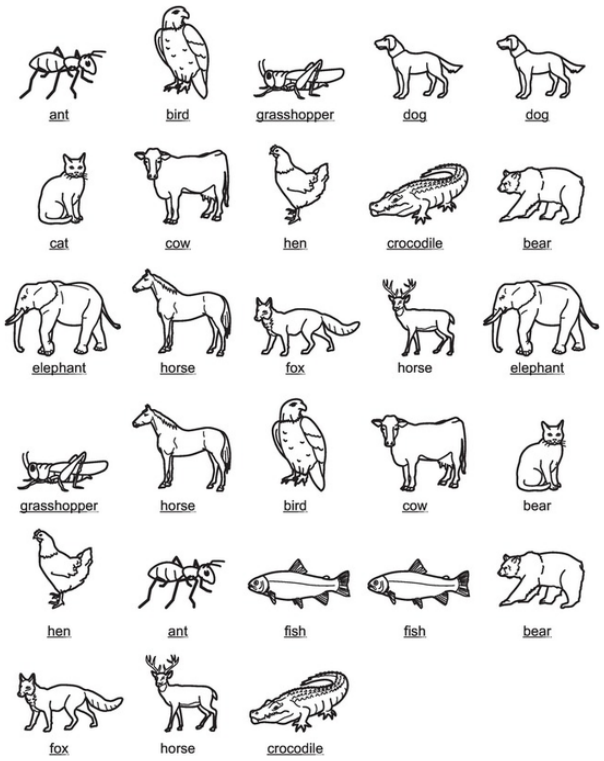


Figure 15.3 D.B.'s responses to pictures of animals presented in his blind field. Correct answers are underlined.

(From Trevethan et al. 2007 .)

The figures were shown at very low contrast (2 percent — although they are depicted in high contrast in Figure 15.3). The patient was told that he was being shown a picture of an animal and asked to guess which animal it was. The figure indicates the responses given, with correct answers underlined. As illustrated, D.B. achieved 89 percent accuracy, despite reporting no awareness whatsoever of any of the figures.

Patients with spatial neglect also have considerable residual abilities. A famous example identified by neuropsychologists John Marshall and Peter Halligan is illustrated in Figure 15.4 . Marshall and Halligan showed P.S., a patient with neglect, the two pictures in the diagram — one of a normal house and one of a house on fire. Because the flames were on the left-hand side of the picture, P.S. did not report seeing any difference between the two pictures. Nonetheless, when asked which house she would prefer to live in, P.S. reliably chose the house that was not on fire (nine times out of eleven).



Figure 15.4 An illustration of the two houses presented to P.S. The houses are identical, except that one has flames shooting out of its left side. Because P.S. possesses left-side spatial neglect, she reported not being able to see the flames but still consistently selected the other house when asked which house she would prefer to live in.

(From Marshall and Halligan 1988 .)

Italian neuropsychologists Anna Berti and Giacomo Rizzolatti (1992) used a semantic priming paradigm to test for semantic categorization in the

neglected field. Patients were presented with priming stimuli in their neglected visual field and then asked to categorize objects in the normal field. As discussed previously, the assumption for priming experiments is that when the prime stimulus and the target stimulus are congruent (i.e., from the same category), then categorization will be easier and quicker, provided that the prime stimulus is processed. Berti and Rizzolatti found the predicted effect in patients who denied all awareness of the prime stimuli and so concluded that semantic information is processed in the neglected visual field .

15.2 So What Is Consciousness For?

Experiments on subjects with brain damage and normal subjects indicate that many information-processing tasks can be performed without conscious awareness — or, more precisely, can be performed by subjects who do not report any conscious awareness of the discriminations and selections that they are making. This leaves us with a puzzle: What exactly does consciousness contribute? Why do we need it? In order to make progress on this, we need to look not just at what patients with blindsight and neglect can do but also at what they can ′ t do.

What Is Missing in Blindsight and Spatial Neglect

Patients with blindsight and neglect have considerable residual abilities that can be teased out with careful experiments, but there are still massive differences between these patients and normal subjects.

Neither patients with blindsight nor those with neglect will *voluntarily* do things in their blind or neglected fields. This is most obvious in patients with neglect. These patients do not just report a complete lack of awareness of what is going on in the neglected visual field. They also do not direct any actions within those regions of space that fall within the neglected visual field. This is the case both for their own personal, bodily space (so that male patients do not shave on the neglected side of their face) and for external space (so that they do not direct actions at objects located on the neglected side of the world as they perceive it). The same holds for patients with blindsight, who never initiate actions toward the blind field, despite being able to point to stimuli in the blind field (when forced to do so).

This suggests a hypothesis about the difference between conscious and nonconscious information processing. Both normal subjects and those with brain damage receive many different types of nonconscious information about the world and about their own bodies. But subjects can only initiate voluntary actions on the basis of information that is conscious. Only conscious information allows subjects to identify targets and plan actions toward them. Neuropsychologists David Milner and Melvyn Goodale have developed a sophisticated theory of vision around the thesis that one of the roles of consciousness is to permit voluntary and deliberate action.

Milner and Goodale: Vision for Action and Vision for Perception

Milner and Goodale ' s theory is based on the existence of two anatomical pathways carrying visual information in the primate brain. As we saw in Section 3.2 , visual information takes two different routes from the primary visual cortex. One pathway, the ventral pathway, projects to the temporal

lobe. A second pathway, the dorsal pathway, carries information to the posterior parietal lobe. (See Figure 3.5 for an illustration of the two pathways.)

The two pathways have very different functions. For Mishkin and Ungerleider, as we saw in Chapter 3 , the crucial functional distinction is between the " what " system, concerned with object identification and subserved by the ventral pathway, and the " where " system, concerned with locating objects in space. Milner and Goodale have a related but somewhat different interpretation. They distinguish two types of vision, which they term *vision for action* and *vision for perception.*

Vision for action has to do with how actions are executed. When you reach out to grasp something, for example, your hand automatically prepares itself so that the fingers are at an appropriate aperture. This involves complex information processing, including estimates of the likely size of the object, taking into account distance and so forth. We all also constantly engage in online correction of movement, compensating for environmental change or initial errors of trajectory. The relevant processing here takes place in the dorsal pathway.

Vision for perception, on the other hand, deals with actually initiating deliberate action. Vision for perception allows targets to be identified and goals to be set. It depends on information from the ventral stream. According to Goodale and Milner, only information relevant to vision for action is actually conscious. Conscious awareness is restricted to the ventral pathway while the dorsal stream governs the visual control of movement nonconsciously.

Milner and Goodale rely heavily on experimental studies of both normal subjects and those with brain damage. Here are two examples that

illustrate how consciousness is and is not involved in vision.

Milner and Goodale ′ s patient D.F. is one of the most studied and important neuropsychological patients. After carbon monoxide inhalation, D.F. developed *visual form agnosia* , substantially impaired visual perception of shape and orientation. The cause was very serious damage to the ventral pathway (i.e., the system responsible for consciously selecting goals and initiating action).

D.F. performs many visuomotor tasks successfully, even though she cannot recognize or identify the relevant features in her environment, as demonstrated by the tasks in Figure 15.5 . When asked to " post " a card into a slot, D.F. was able to match her movements to the orientation of the slot and performed almost as successfully as normal subjects. But when asked to make an explicit judgment about the slot ′ s orientation, D.F. ′ s responses were almost random. This was the case whether she was asked to describe the orientation verbally or nonverbally (by rotating a card to match the orientation). According to Milner and Goodale, D.F. is receiving nonconscious information about orientation through the dorsal pathway, but because of damage to her ventral pathway, she is not consciously aware of the orientation.

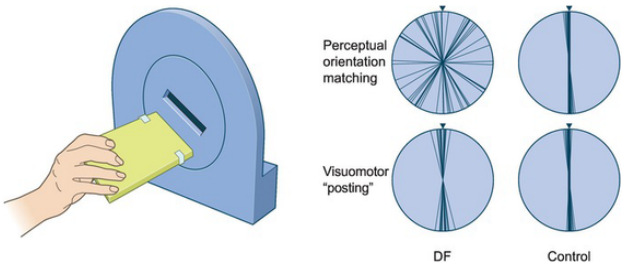


Figure 15.5 In this experiment, subjects were asked either to " post " a card into a slot or to rotate another handheld card to match the orientation of the slot. The angle of the slot varied across trials, although in each case, the diagrams were normalized so that the correct result was vertical. Normal subjects were able to perform both tasks with little difficulty. Patient D.F., in contrast, carried out the visuomotor task almost as well as normal subjects, but her responses in the explicit matching task were almost random.

(From Milner and Goodale 1998 .)

Visual illusions also show a dissociation between (nonconscious) vision for action and (conscious) vision for perception. Visual illusions affect how subjects consciously perceive the size and shape of objects. Experimenters have found, however, that the illusion does not carry over to visuomotor behavior. Subjects will report seeing an illusion, but when asked to make appropriate movements, they will configure their grip and make other adjustments according to the correct dimensions of the relevant objects, not the dimensions that they report perceiving. So conscious perception (vision for perception) dissociates from (nonconscious) information relevant to the control of visuomotor behavior (vision for action). Figure 15.6 illustrates the experiment used by Aglioti, DeSouza, and Goodale to identify this dissociation, utilizing the *Ebbinghaus illusion* .

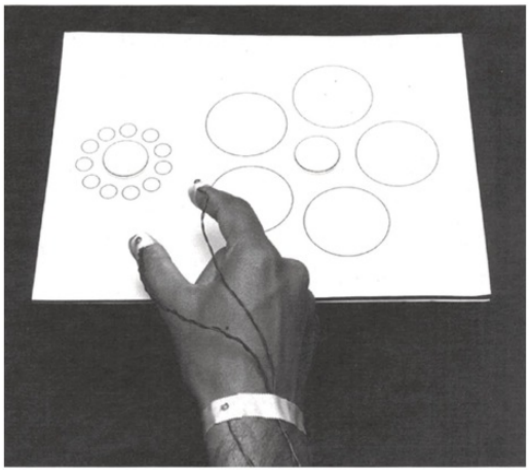


Figure 15.6 In the Ebbinghaus illusion, two circles are illusorily seen as differently sized, depending on what surrounds them. The figure illustrates experiments published by Aglioti et al. in 1995. The experimenters measured the size of the opening between the fingers and thumb when subjects were asked to pick up two disks that they reported as being differently sized. They found no significant differences in grip aperture, suggesting that this aspect of the fine-grained control of grasping draws on different types of visual information than those that yield conscious awareness of the disks.

(From Aglioti et al. 1995 .)

In addition, neuroimaging studies, such as those published by Fang Fang and Shen He in 2005, suggest that ventral stream activity is correlated with consciousness, whereas activity in the dorsal stream is not. Fang and He compared activation levels in areas known to be very involved in object processing in the dorsal and ventral streams, respectively. They used a technique known as *interocular suppression* in which one eye is presented with an image of an object while the other eye is presented simultaneously

with a high-contrast pattern that blocks conscious awareness of the presented image.

This paradigm enabled Fang and He to examine activation levels in the dorsal and ventral streams in the absence of conscious awareness and compare those levels with activation levels when conscious awareness of the image was not suppressed. They found robust levels of activity in the dorsal stream even in the nonconscious conditions. In contrast, ventral stream activation was confined to the conscious condition.

In conclusion, Milner and Goodale ' s distinction between (conscious) vision for perception and (nonconscious) vision for action, together with the evidence supporting it from normal subjects and those with brain damage, both supports and clarifies the hypothesis that consciousness is important for initiating action. If Milner and Goodale are correct, then conscious awareness is key for identifying targets and for macro-level planning for how to effect actions. But conscious awareness is not typically involved in the fine-grained, online control of bodily movements .

What Is Missing in Masked Priming

Although masked priming experiments offer powerful evidence of nonconscious semantic processing, this is very different from normal, conscious information processing.

One key finding is it is hard to retain information without conscious awareness. Semantic information processed below the threshold of consciousness in masked priming experiments is very transitory and short-lived. Here is an illustration from experiments published by Anthony Greenwald, Sean Draine, and Richard Abrams in 1996. The authors used a

typical categorization task, asking subjects to identify first names as male or female or to classify words as pleasant or unpleasant in meaning. They succeeded in eliciting a robust priming effect when subjects were presented with a congruent masked prime. This effect was present both when the prime was presented subliminally and when it was presented supraliminally (above the threshold of consciousness). This allowed Greenwald et al. to study the differences between subliminal priming and supraliminal priming by exploring what happened when they varied the time between prime and trial (the *stimulus-onset asynchrony* [SOA]).

Greenwald et al. found that in supraliminal cases, where the subjects were conscious of the prime, the priming effect was robust across all SOAs. The length of the delay between prime and target did not make a significant difference. In contrast, in the subliminal cases, with the subjects not consciously perceiving the prime, the effect was robust only at the shortest intervals and disappeared completely once the SOA went above 100 ms.

This experiment suggests an additional hypothesis about the function of conscious awareness, namely, that consciousness allows information to be explicitly retained and maintained. According to this hypothesis, information that is picked up nonconsciously can indeed be deployed in relatively sophisticated tasks, but it can be used only within a very limited time horizon. Conscious information, in contrast, is more transferable and flexible. It can be used beyond the here and now. There are definite parallels between this idea and Goodale and Milner ' s distinction. Vision for action is restricted to the online control and fine-tuning of behavior. It does not persist in the way that conscious visual information persists. That is one reason why Goodale and Milner think that the conscious vision-for-perception system is required for high-level action planning.

Exercise 15.4 Summarize in your own words the two hypotheses that have emerged about the nature of consciousness, and sketch out the evidence for them.

15.3 Two Types of Consciousness and the Hard Problem

Philosopher Ned Block has cautioned cognitive scientists against drawing sweeping conclusions about consciousness from neuropsychological disorders such as blindsight and unilateral spatial neglect. He thinks that these conclusions rest on a confusion between two very different concepts of consciousness. He calls these *phenomenal consciousness* and *access consciousness*. Here is how he characterizes them:

Phenomenal consciousness

> P-consciousness is experience. . . . We have P-conscious states when we see, hear, smell, taste, and have pains. P-conscious properties include the experiential properties of sensations, feelings, and perceptions, but I would also include thoughts, wants, and emotions.
>
> (Block 1995b : 230)

Access consciousness

> A state is A-conscious if it is poised for direct control of thought and action. To add more detail, a representation is A-conscious if it is poised for free use in reasoning and for direct " rational " control of

action and speech. (The rational is meant to rule out the kind of control that obtains in blindsight.)

(Block 1995b : 231)

Exercise 15.5 Give your own examples of A-consciousness and P-consciousness, and describe the difference between them in your own words.

For Block, the real problem of consciousness is the problem of understanding P-consciousness. All of the things that we have been looking at in the previous section, however, are really examples of A-consciousness. They do not get to the heart of what Block sees as the real problem of consciousness, which has to do with how and why we experience the world the way we do.

Block ′ s distinction between A-consciousness and P-consciousness is related to influential ideas developed by philosopher David Chalmers in his 1996 book *The Conscious Mind* . Chalmers thinks that there is no single problem of consciousness. Instead, we need to make a distinction between a cluster of relatively easy problems and a single, really difficult problem — what he calls the " hard problem " of consciousness.

Here are some examples of what Chalmers provocatively identifies as easy problems:

- Explaining an organism ′ s ability to discriminate, categorize, and react to environmental stimuli
- Explaining how a cognitive system integrates information

- Explaining how and why mental states are reportable
- Explaining how a cognitive system can access its own internal states
- Explaining how attention gets focused
- Explaining the deliberate control of behavior
- Explaining the difference between wakefulness and sleep

In Block ' s terminology, these are different aspects of understanding A-consciousness. They all have to do with how an organism accesses and deploys information.

Chalmers recognizes that " easy " is a relative term. None of the so-called easy problems has yet been solved — or even partially solved. The reason he calls them easy problems is that at least we have some idea of what a solution would look like. The easy problems are all problems that are recognizable within the basic framework of cognitive science and scientific psychology. People write papers about them, reporting relevant experiments and constructing theories.

According to Chalmers, though, no amount of progress on the easy problems of consciousness will help with the hard problem. Here is how he characterizes the hard problem:

> The really hard problem of consciousness is the problem of *experience.* When we think and perceive, there is a whir of information-processing, but there is also a subjective aspect. As Nagel (1974) has put it, there is *something it is like* to be a conscious organism. This subjective aspect is experience. When we see, for example, we *experience* visual sensations: the felt quality of redness, the experience of dark and light, the quality of depth in a visual field.

> Other experiences go along with perception in different modalities: the sound of a clarinet, the smell of mothballs. Then there are bodily sensations, from pains to orgasms; mental images that are conjured up internally; the felt quality of emotion, and the experience of a stream of conscious thought. What unites all of these states is that there is something it is like to be in them. All of them are states of experience.
>
> (Chalmers 1995 : 201)

It is undeniable that some organisms are subjects of experience. But the question of how it is that these systems are subjects of experience is perplexing. Why is it that when our cognitive systems engage in visual and auditory information processing, we have visual or auditory experience: the quality of deep blue, the sensation of middle C? How can we explain why there is something it is like to entertain a mental image, or to experience an emotion?

If any problem qualifies as *the* problem of consciousness, it is this one.

Exercise 15.6 In your own words, characterize the proposed distinction between what Chalmers calls the easy problems of consciousness and what he calls the hard problem .

We can bring this idea that there is a hard problem of consciousness into clearer focus through the Knowledge Argument, a thought experiment originally proposed by the philosopher Frank Jackson. Here is how he sets it up:

> Mary is confined to a black-and-white room, is educated through black-and-white books and through lectures relayed on black-and-white television. In this way she knows everything there is to know about the physical nature of the world. She knows all the physical facts about us and our environment, in a wide sense of " physical " which includes everything in *completed* physics, chemistry, and neurophysiology.
>
> (Jackson 1986 : 291)

Now imagine that Mary emerges from the black-and-white room and sees a red rose for the first time. Nothing that she learned in the black-and-white room will have prepared her for the experience, even though she has a complete scientific understanding of herself and of the world (including, of course, a complete information-processing account of color perception). So it follows, according to Jackson, that information-processing models leave something out. They cannot explain the nature of conscious experience.

Exercise 15.7 Formulate Jackson ' s Knowledge Argument in your own words. How convincing do you find it?

The Knowledge Argument is a good way of capturing Chalmers ' s basic idea that the *hard* problem of *phenomenal consciousness* is, in principle, beyond the reach of cognitive science. All cognitive science can do, in this view, is tackle the *easy* problem of *access* consciousness. But should we accept this way of looking at things?

Without trying to settle the matter one way or the other, it seems plausible that progress is going to depend on having a better idea of what an information-processing account of access consciousness might look like. Discussing the limits that there might or might not be to a particular type of explanation will be much easier when there is a particular example on which to focus. The next section will consider the two leading theoretical approaches.

15.4 Theories of Consciousness

Higher-Order Theories

Higher-order theories of consciousness have been developed primarily within philosophy, but the approach has been influential within cognitive science more generally. Prominent exponents include Peter Carruthers, William Lycan, David Rosenthal, and the late David Armstrong.

There are many different types of higher-order theories, but they all offer answers to a single basic question: How can we explain the difference between mental states that are conscious and those that are not? The basic idea behind higher-order approaches is deceptively simple. Mental states are conscious when we are aware of them, and they are nonconscious when we are not aware of them. Higher-order thought theories hold that this awareness comes from a higher-order state that is directed at the original, first-order state. A first-order state, whether a thought, a perception, or an intention, is conscious when it is the object of a higher-order mental state.

This way of thinking about consciousness is consistent with much (but, as we will see, not all) of the evidence reviewed earlier. For example, it

certainly allows for nonconscious information processing, which will employ mental states that are not the objects of any sort of higher-order awareness. In fact, according to higher-order theories, the very same state can be conscious at some times and nonconscious at others, depending on whether it is the object of a higher-order state.

The principal division among higher-order theories lies in how they understand the higher-order states that confer consciousness on lower-order states.

According to higher-order perception (HOP) theories, developed by Armstrong and Lycan, first-order states are conscious when they are the objects of what is sometimes called an *inner sense* . Armstrong and Lycan understand inner sense as a form of what is traditionally known as *introspection* , whereas other proponents take it to be a quasi-perceptual faculty that " scans " first-order representations to generate sensory representations of them.

Each form of HOP theory faces difficulties. " Introspection " is hardly a clear notion, and the most natural way of defining it would be as the tool or mechanism by which we are aware of our own mental states. But *aware* is in many ways a synonym for *conscious* , and thus that form of the HOP theory quickly starts to look circular. The second form is more promising as an account of the paradigm cases of what Block calls *phenomenal consciousness* , such as perceptual " feels " and other directly sensory experiences (pains, itches, etc.). It is rather implausible, however, as an account of conscious beliefs and thinking more generally. In what sense can I be said to have a sensory representation of my thought that the square of a triangle ' s hypotenuse is equal to the sum of the squares of the other two sides?

Higher-order thought (HOT) theories avoid these difficulties by holding that the higher-order state that confers consciousness is a thought, not a perception. Just as we can become aware of a table by thinking about it, the HOT theorist suggests that we become aware of mental states by thinking about them. According to Rosenthal, for example, mental states are conscious when, and only when, they are accompanied by a suitable HOT with that mental state as its object.

HOT theories face challenges of their own. One objection frequently raised is that it remains mysterious how thinking about something can make it conscious. My table does not become conscious when I think about it, so why should things be different for my thought that the square of a triangle ′ s hypotenuse is equal to the sum of the squares of the other two sides? A HOT theorist cannot reply that the HOT is itself conscious because that would lead to either circularity or regress. Moreover, and this is also a problem for HOP theories, the higher-order approach has seemed to many to misrepresent the structure of consciousness. When I am consciously smelling a rose, what I am aware of is the rose and its smell, not a mental state, as both the HOP and HOT theories seem to imply.

HOT theorists have a range of sophisticated responses to these and other objections. Moreover, they cite various types of empirical support for their approach to consciousness. In particular, they point to studies showing that conscious awareness is correlated with activation in the prefrontal and parietal cortices, areas generally thought to be involved in complex cognitive behaviors and control. In one set of studies carried out by Hakwan Lau and Richard Passingham, for example, visual masking was used to vary subjective awareness in a visual discrimination task. They found that performance on the task remained constant through changes in subjective

awareness (as measured by subjects ' reports of whether or not they were guessing). Activation levels in the dorsolateral prefrontal cortex (Brodmann area 46) were correlated with reported awareness.

In one important respect, however, HOT theories do not sit well with the empirical evidence discussed earlier. It is hard to see how the distinctive functional role of consciousness can be explained by the presence/absence of HOTs because it seems to be part of the HOT theory that mental states behave exactly the same way whether or not they are the objects of HOTs. In fact, empirically sophisticated HOT theorists point to studies like that of Lau and Passingham, in which conscious awareness seems not to make much difference. But still, HOT theorists have an uphill struggle in explaining away all empirical evidence suggesting that consciousness has a distinct functional role.

The Global Workspace Theory

In contrast to higher-order theories, global workspace theory is firmly focused on the functional role of consciousness. It was originally proposed by psychologist and cognitive scientist Bernard Baars in his book *A Cognitive Theory of Consciousness* , published in 1988. Since then, it has been taken up and developed by many others, including neuroscientists Antonio Damasio and Stanislas Dehaene.

In their 2001 article " Towards a Cognitive Neuroscience of Consciousness, " Dehaene and Naccache present the theory as explaining three central functions of consciousness. These are as follows:

- The intentional control of action
- Durable and explicit information maintenance

- The ability to plan new tasks through combining mental operations in novel ways

The global workspace theory is basically a hypothesis about mental architecture and the large-scale organization of the mind that explains how these functions can be carried out.

Dehaene and Naccache emphasize the distinction between modular processing and nonmodular processing. Modular information processing is domain specific and encapsulated. It is also, they think, nonconscious. Only information processing that is domain general and integrates multiple different sources can be conscious. This is the basic idea behind the metaphor of the global workspace.

> **Exercise 15.8** Review the discussion of modularity in Chapter 8 .

For the global workspace theory, attention and consciousness are very closely linked. Attention functions as a gatekeeper, allowing the results of modular information processing to enter the global workspace. This way of thinking about the role of attention has a long pedigree within cognitive science, going back to the pioneering work of Donald Broadbent, reviewed in Section 1.4 . Attention is thought of both as a *filter* (screening out unnecessary information, as in the cocktail party effect) and as an *amplifier* (allowing information that would otherwise have been unconscious to become available to consciousness).

The Global Neuronal Workspace Theory

Three versions of the global workspace theory are illustrated in Figure 15.7 , showing how the workspace idea has evolved over the last thirty years.

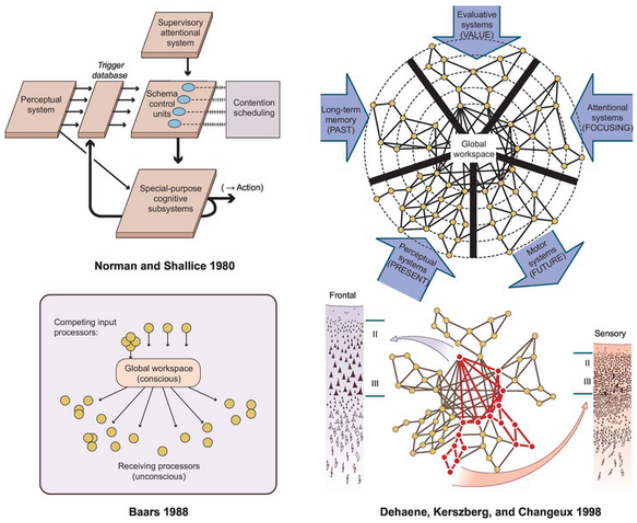


Figure 15.7 In the Norman and Shallice 1980 model (*top left*), conscious processing is involved in the supervisory attentional regulation, by prefrontal cortices, of lower-level sensorimotor chains. According to Baars 1988 , conscious access occurs once information gains access to a global workspace (*bottom left*), which broadcasts it to many other processors. The global neuronal workspace (GNW) hypothesis (*right*) proposes that associative perceptual, motor, attention, memory, and value areas interconnect to form a higher-level unified space where information is broadly shared and broadcasted back to lower-level processors. The GNW is characterized by its massive connectivity, made possible by thick layers II/III with large pyramidal cells sending long-distance cortico-cortical axons, particularly dense in prefrontal cortex.

(From Dehaene and Changeux 2011 .)

The first version (under a different name) was the theory of attention originally proposed by Donald Norman and Tim Shallice. As the figure illustrates, attention performs what Norman and Shallice term *contention scheduling*. Contention scheduling is required when different cognitive systems propose competing responses (whether cognitive or behavioral) to a single set of stimuli. Contention scheduling effectively resolves the competition to select a single response, which can either be an output to the action systems or can be fed back into the cognitive systems.

The terminology of *global workspace* was introduced by Bernard Baars in the late 1980s. One version of his theory is depicted in the figure, showing very clearly how the global workspace is envisioned as a conscious window between nonconscious inputs and conscious outputs.

A much more recent version of the theory is depicted on the right side of Figure 15.7 . It was developed by Stanislas Dehaene, Michel Kerszberg, and Jean-Pierre Changeux (1998). This shares some features with the other two versions, particularly the idea that the global workspace receives inputs from different cognitive modules and then sends outputs to motor systems. The Dehaene et al. theory is much more strongly grounded in hypotheses about neural implementation and connectivity — which is why they call their theoretical construct the *global neuronal workspace* , rather than simply the *global workspace* . This emerges even more clearly in Figure 15.8 .

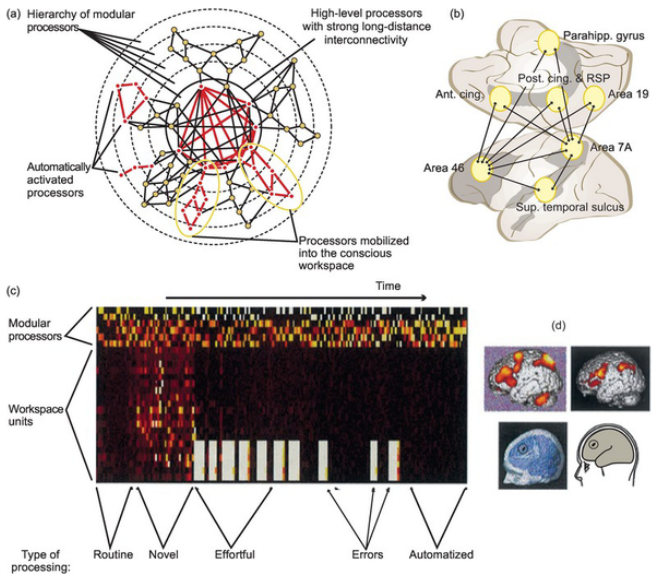


Figure 15.8 The neural substrates of the global workspace. (a) Hierarchy of connections between different processors in the brain. Note the strong long-distance connections possessed by the higher levels. (b) Proposed anatomical substrate of the global workspace. This includes a network linking the dorsolateral prefrontal, parietal, temporal, and anterior cingulate areas with other subcortical regions (RSP = retrosplenial region). (c) Neural dynamics of the global workspace, derived from a neural simulation of the model shown in panel a. The activation levels of various processor units (*top lines*) and workspace units (*bottom lines*) are shown as a function of time. (d) Different parts of the global workspace network activated by different tasks, including the generation of a novel sequence of random numbers, effortful arithmetic, and error processing.

(From Dehaene and Naccache 2001 .)

Figure 15.8 makes plain the distributed nature of the global neuronal workspace, as envisaged by Dehaene and his collaborators. They see the modular part of the mind as composed of many interconnecting modules

that feed into each other in a hierarchical manner. (The hierarchy is depicted by concentric circles, and the closer the circles to the center, the higher their place in the hierarchy.) Some of the hierarchical modules form fully automatic and nonconscious networks. Others, in contrast, have amplified levels of activity that allow them to feed into the global workspace.

The global neuronal workspace itself is not a single neural location (as the metaphor of a workspace might initially suggest) but, rather, a distributed network of high-level processors that are highly connected to other high-level processes. The candidate areas identified include the prefrontal, parieto-temporal, and cingulate cortices — all areas that we have discussed in the context of different types of high-level cognition at various points in this book. The lower portion of Figure 15.8 includes a neural network simulation of the global neuronal workspace and when it becomes engaged, in addition to a functional magnetic resonance imaging (fMRI) diagram indicating activation levels across the hypothesized network during high-level conscious tasks such as mental arithmetic.

The hypothesis is that the global neuronal workspace is generated by the activities of a particular type of neuron called a *pyramidal neuron* . Pyramidal neurons are very widespread in the mammalian cortex and particularly dense in the prefrontal, cingulate, and parietal regions (all hypothesized to be important in the global neuronal workspace). They are characterized by a single long axon and heavily branched dendrites, which allow them to communicate with many other neurons and with distant brain areas. Dehaene and collaborators hypothesize that networks of pyramidal neurons connect specialized modular processes and allow their outputs to be broadcast across the brain so that they are available for action planning, verbal report, and other high-level cognitive processes.

Exercise 15.9 Explain in your own words how the global neuronal workspace theory incorporates the hypotheses about the function of consciousness identified in Section 14.4 .

15.5 Conclusion

There are two very different approaches to consciousness. On the one hand, there are those who think that consciousness is a mystery that we have no idea how to tackle with the tools and methods of cognitive science. On the other hand, we have thriving research programs that study different aspects of the conscious mind and how consciousness contributes to action and cognition.

The " mysterians, " as they are sometimes called, hold that the various research programs we have looked at only touch upon the " easy " aspects of the problem of consciousness — at best, they can only explain access consciousness, as opposed to the really tough problem of explaining how and why we are phenomenally conscious. The global neuronal workspace theory, for example, was primarily developed to explain how consciousness can make a difference to cognition. The theory gives an account of why some information becomes conscious and how that information has a distinctive role to play in higher-level cognition. Mysterians will say that this account is all well and good, but it cannot come to grips with the " hard problem " of explaining the distinctive experience of being conscious.

In response, cognitive scientists working on consciousness may well respond that the so-called *hard problem* of consciousness will disappear

once we have a good enough understanding of the various phenomena lumped together under the label " access consciousness. " This is the view taken by the philosopher Daniel Dennett, whose books *Content and Consciousness* (1969) and *Consciousness Explained* (1991) have been very influential in discussions of consciousness.

More generally, there is a lesson to be learned from debates in the nineteenth and early twentieth centuries about vitalism in biology. Vitalists such as philosopher Henri Bergson and biologist John Scott Haldane believed that the mechanist tools of biology and chemistry were, in principle, incapable of explaining the difference between living organisms and the rest of the natural world. Instead, we need to posit a vital force, or *élan vital* , that explains the distinctive organization, development, and behavior of living things.

Certainly, vitalism has no scientific credibility today. The more that was discovered about the biology and chemistry of living things, the less work there was for an *élan vital* , until finally it became apparent that it was an unnecessary posit because there was no problem to which it might be a solution. But historians of science argue that debates about vitalism served an important role in the development of biology by forcing biologists to confront some of the explanatory deficiencies of the models they were working with — both by developing new models and by developing new experimental tools. Perhaps mysterianism about the cognitive science of consciousness will have a similar role to play?

Certainly, that would be consistent with how cognitive science has evolved up to now. Many of the advances that we have explored have emerged in response to challenges that, on the face of things, are no less dramatic than the challenges posed by those who think that consciousness is

scientifically inexplicable — the challenge to show how a machine can solve problems, for example; to show how neural networks can learn; or to show how systems can engage in sophisticated emergent behaviors without explicit information processing.

In any event, consciousness is one of the most active and exciting topics in contemporary cognitive science. Whether it will ultimately reveal the limits of cognitive scientific explanation or not, it continues to generate an enormous range of innovative experiments and creative theorizing.

Summary

This chapter explored the basic challenge for cognitive science raised by the apparent conflict between first- and third-person approaches to consciousness. We began with results from priming studies and neurological damage indicating that a great deal of information processing occurs below the threshold of consciousness. Further investigation shows, though, that nonconscious information processing is limited in important ways, and we looked at those limitations to explore the function of consciousness awareness. Milner and Goodale ′s research on the two visual streams, as well as other related studies, indicate that consciousness is important for planning and initiating actions, whereas research on planning suggests that memory retains conscious information better than nonconscious information. After considering objections to the very idea that cognitive science might provide a complete account of consciousness, we concluded by looking at two prominent contemporary theories of

consciousness — the higher-order thought theory and the global workspace theory.

Checklists

Information Processing without Conscious Awareness

1. There are two primary ways of understanding unconscious information processing: priming experiments and studies of patients with neurological damage.

2. Semantic priming studies show that basic categorization can be accomplished unconsciously. Because semantic categorization is generally thought to be nonmodular, these tasks also suggest that there can be nonmodular unconscious processing.

3. Blindsight and unilateral neglect indicate that high-level processing can be applied even to areas of the visual field that, as a result of damage, are not consciously perceived.

The Function of Consciousness

1. Milner and Goodale ′s research reveals a basic functional distinction in the visual system: vision for perception and vision for action. The ventral visual stream is for perception and high-level action planning and is conscious, whereas the dorsal visual stream is for online control of action and is unconscious.

2. Experiments on the Ebbinghaus illusion and interocular suppression provide support for Milner and Goodale ′s dual-stream hypothesis.

3. Milner and Goodale ′s research indicates that consciousness is functionally important for planning and initiating action.

4. Priming studies show that consciously perceived primes are retained better and have a greater impact on other cognitive processes.

The Challenge of Consciousness

1. Ned Block ′s distinction between access consciousness (or A-consciousness) and phenomenal consciousness (or P-consciousness) can be used to generate a dilemma for the cognitive science of consciousness. Cognitive science seems to be informative only for understanding A-consciousness, but P-consciousness is what really matters to us.

2. Block claims that there is a double dissociation between A- and P-consciousness. This produces an explanatory gap.

3. The conflict between A- and P-consciousness can be understood in terms of what David Chalmers calls the *hard problem of consciousness* , which he thinks cognitive science cannot solve.

4. Jackson ′s Knowledge Argument reaches a similar conclusion.

5. According to the Knowledge Argument, a colorblind scientist who knew every (third-person) scientific fact about consciousness would still lack a crucial piece of knowledge — first-person knowledge of what it is like to see color.

The Higher-Order Thought Theory of Consciousness

1. According to higher-order theories, mental states are conscious when we are aware of them.

2. Higher-order perception (HOP) theories hold that mental states are conscious when they are the objects of inner sense (Armstrong) or "scanned" by a quasi-perceptual faculty (Lycan).

3. For higher-order thought (HOT) theories, what confers consciousness is a thought, not a perception.

4. Both HOP and HOT theories have been objected to (a) for obscuring the fact that we are aware of things in the world, not our own mental states, and (b) for not doing justice to the distinctive functional role of consciousness.

The Global Workspace Theory of Consciousness

1. The global workspace theory holds that attention makes low-level modular information available for conscious control in the "global workspace," from where the information is then "broadcast" to other areas of the brain.

2. The global workspace theory draws from two basic ideas: (a) consciousness permits information to be explicitly and durably maintained for additional processing and reasoning, and (b) consciousness is necessary for initiating deliberate action.

3. Information processing in the global workspace is a type of domain-general process, selecting among competing modular inputs.

4. There is some neurological support for the global workspace theory. Pyramidal neurons, for instance, may be responsible for connecting specialized modular processes and broadcasting their outputs throughout the brain for other cognitive processes.

Further Reading

There has been an explosion of research on consciousness in the last decade or so, only a small portion of which can be covered in a single chapter. Good places to start to learn more are the books by Jesse Prinz 2012 and Timothy Bayne 2012 . Although written by philosophers, both books place heavy emphasis on empirical research and synthesize a wide range of recent studies of consciousness. Robert Van Gulick ′ s chapter in Margolis et al. 2012 also provides a good summary of both philosophical and neuroscientific theories of consciousness. Zelazo et al. 2007 is another excellent resource. Baars and Gage 2010 discusses a lot of the most recent research, including figures and descriptions of the most popular methods used to study consciousness. Chalmers 2013 outlines what he sees as the principal projects for a science of consciousness.

Frank Jackson ′ s Knowledge Argument was first presented in Jackson 1982 . His more recent views can be found in Jackson 2003 . Ludlow et al. 2004 and Coleman 2019 are collections of essays on the Mary thought experiment. See also Nida-Rümelin and Conaill 2021 in the online *Stanford Encyclopedia of Philosophy* (https://plato.stanford.edu/). For a related argument (Joseph Levine ′ s explanatory gap argument), see Levine 1983 .

Prominent accounts of how unconscious information processing operates and how information becomes conscious include Dehaene et al. 2006 and Kouider et al. 2007 . There are many excellent reviews of research on priming. Kouider and Dehaene 2007 is a good survey of the history of masked priming. Recent studies include Avneon and Lamy 2018, 2019 . On primes becoming more visible when followed by congruent primes, see Bernstein et al. 1989 . Good resources on bilingual semantic priming are Kiran and Lebel 2007 , Kotz 2001 , and Schoonbaert et al. 2009 . Classic studies of unilateral neglect include Driver and Mattingly 1998 , Driver and Vuilleumier 2001 , and Peru et al. 1996 . A recent meta-analysis of the critical lesion locations involved in unilateral neglect can be found in Molenberghs et al. 2012 . Corbetta and Shulman 2011 discusses the relation between neglect and attention. On the function of the parietal cortex in visual perception, see Husain and Nachev 2007 . Phillips 2021 argues that blindsight is better viewed as degraded conscious perception.

A summary of the two visual streams can be found in Milner and Goodale 2008 . A critique of the two-stream account (with commentary from Milner, Goodale, and others) can be found in Schenk and McIntosh 2010 . See Milner 2012 for a recent study of the two visual streams and consciousness. Goodale and Milner 2013 also provides a good review of the visual system. There are many studies on the Ebbinghaus illusion and the differences between vision for action and vision for perception. Aglioti et al. 1995 is a classic study. For responses and follow-up studies, see Glover and Dixon 2001 and Franz et al. 2000 .

The literature on access consciousness and phenomenal consciousness is quite large now. Block 1995b is the classic article on the topic. Block ' s more recent views can be found in Block 2007 , where he proposes different

neural structures underlying A- and P-consciousness, and Block 2011 , where he responds to a number of criticisms of his account. The original Sperling experiment can be found in Sperling 1960 . A criticism of Block ' s interpretation of the Sperling experiment, as well as discussion of phenomenal consciousness more generally, can be found in Kouider et al. 2010 . For more on the putative explanatory gap between A- and P-consciousness, see Levine 1983 . Other well-known books on these topics include Carruthers 2000 and Dennett 1991 . For classic formulations of the hard problem and easy problems of consciousness, see Chalmers 1995 , 1996 .

Rosenthal 1986 , 1997 are influential sources for the HOT theory. The later paper is published in the Block et al. 1997 anthology, which also has pieces by Lycan on HOPs and Armstrong on HOTs. Lau and Rosenthal 2011 presents the empirical case for HOT theory. For early formulations of the global workspace theory of consciousness, see Baars 1988 , 2002 . An influential discussion of the theory is Dehaene and Naccache 2001 . A summary of the theory can be found in Dehaene and Changeux 2011 , including responses to critics. See also Dehaene et al. 2014 , Mashour et al. 2020 , and Whyte and Smith 2021 .

Two popular topics in consciousness research that have been mentioned only briefly but that have their own burgeoning literatures are attention, which was discussed in Chapter 9 , and the neural correlates of consciousness. Posner 1980 is a classic early study on attention and consciousness. It was the first to convincingly demonstrate that gaze can be fixed while attention wanders. Lamme 2003 provides a concise summary of the reasons for separating attention from consciousness. Lavie 2005 is an influential account of how unattended stimuli are processed. Mack and

Rock 1998 discusses a series of now-classic experiments on what is called *inattentional blindness* . Simons and Chabris 1999 is another classic series of studies in this area. These experiments rely on selective looking, where people ' s selective attention alters what they see in a visual array. See Simons and Rensink 2005 for a review of these studies. Other reviews of how attention relates to consciousness can be found in Koch and Tsuchiya 2007 , Martens and Wyble 2010 , Van den Bussche et al. 2010, and Raffone et al. 2014 .

Many trace the most recent wave of research into the neural correlates of consciousness (NCCs) to Baars 1988 and Koch 2004 . The global workspace theory is one prominent account of the NCCs. An influential idea utilized by global workspace theorists is that of neural synchrony. This idea, popularized by Singer 1999 , holds that groups of neurons must fire in sync in order to produce consciousness. Womelsdorf et al. 2007 is a more recent paper on this phenomenon. Crick and Koch 2003 is a widely cited review of different problems with the search for NCCs, including arguments against the importance of neural synchrony for consciousness. An increasingly popular tool for identifying the NCCs is to track brain activation in patients during and after being in a vegetative state. Steven Laureys ' s studies are some of the best known. Laureys 2005 is an influential article describing the various brain areas that appear to be deactivated as a result of being in a vegetative state. Owen et al. 2006 and Hohwy 2009 are other important articles. Good reviews on the search for the NCCs include Metzinger 2000 , Lamme 2006 , Tononi and Koch 2008 , and Koch et al. 2016 . Doerig et al. 2019 discuss criteria for evaluating empirical theories of consciousness.

Chapter 16

The Emotions: From Cognitive Science to Affective Science

◈

Coauthored with Dong An

Overview

16.1 Early Theories

Physical Symbol Systems and the Emotions

Ekman and Basic Emotions

16.2 Affective Space and the Affective Scientist's Toolkit

Elements of an Emotional Episode

Genetic Tools in Affective Science

Lesion Studies

16.3 Fear: A Multilevel and Multidisciplinary Case Study

Fear Conditioning and Fear Responses in Rodents

Fear and Amygdala Damage in Humans

Neuroimaging of Fear in Humans

Overview

This book has so far focused primarily on cognition. This chapter turns to emotional and affective life more generally. This aspect of the mind was somewhat neglected by cognitive scientists in the early decades of the interdisciplinary study of the mind but has received much more attention in recent years. In fact, researchers in the area are increasingly using the phrase " affective science " to describe their enterprise, which is an exciting new direction for cognitive science more generally.

Section 16.1 discusses two early theories. Herbert Simon, a pioneer of the physical symbol system hypothesis, developed a computational approach to emotions as interrupt mechanisms within the central nervous system. From a very different perspective, Paul Ekman proposed that there are a small number of basic emotions that are constant across cultures, have distinct neural bases, and evolved to serve specific functions. Juxtaposing Simon and Ekman ' s theories frames the key questions that affective scientists tackle.

Section 16.2 explores how many affective scientists have moved toward thinking of emotions in multidimensional terms (as opposed to a small number of discrete, basic emotions). It then outlines some of the tools in the affective scientist ' s toolkit. Some, such as functional neuroimaging, should be very familiar from earlier chapters. Others, such as the study of patients with brain damage, are less familiar. And one of the most exciting things about affective science is that it has adopted a suite of new tools, making particular use of new developments in gene-editing technology.

Section 16.3 shows how the affective scientist ' s toolkit can be put to work by focusing on fear, which is probably the most studied emotion. The

studies reviewed in this section include work on fear conditioning in rodents (as well as in humans), influential discoveries from patients with brain damage, and more recent neuroimaging studies. As with other areas of cognitive science, the study of brain function (and dysfunction) plays a key role in affective science.

16.1 Early Theories

Physical Symbol Systems and the Emotions

As we have seen, the early history of cognitive science was dominated by the physical symbol hypothesis and the computer model of the mind. Herbert Simon was a key player in both of these, but it did not escape his notice that the computer model of the mind, as initially formulated, did not have much to say about emotions and affective life in general. In an important paper in 1967 entitled " Motivation and Emotional Controls of Cognition, " Simon set out to show that sufficiently complex serial information processors, which is what he took minds to be, could be expected to incorporate emotion-like mechanisms. In effect, he gives a computational theory of emotions.

Simon thinks that the mind is a hierarchically organized information-processing system that manipulates symbols serially. So, it has a hierarchy of increasingly specific subroutines under the control of an executive program at the highest level. This process of delegation can only work, however, if there exist mechanisms for indicating when subroutines should be terminated. Even the most single-minded system needs goal-completion mechanisms that can do the following:

- Identify when the goal has been completely achieved
- Determine that the goal has been " satisficed " (achieved " well enough " in the circumstances)
- Select the best alternative achieved after a specified amount of time/processing
- Give up when the system has been unsuccessful for a specified amount of time/processing

> **Exercise 16.1** Look back at the account of the General Problem Solver in Chapter 4 and identify a goal-completion mechanism.

What happens with serial systems that have multiple goals? Simon identifies two relatively simple mechanisms that can easily be built into information-processing systems. The first is a mechanism for queuing goals. The second is developing multifaceted criteria for evaluating goal completion. He gives the example of a chess computer, which might evaluate not just when the game is won but also how many pieces have been taken, how much of the board is controlled by each player, who is in command of the center, and so forth.

But these additional mechanisms fail to accommodate the fact that the goals of living organisms are time constrained. Danger needs to be avoided immediately. Bodily needs have to be met on a daily basis. For many species, reproduction is seasonal, and so on. Simon argues, plausibly, that these multiple time constraints can only be integrated by systems that have what he calls *interrupt mechanisms* . Such mechanisms must be capable of

constantly monitoring the environment (and perhaps the system itself) in order to identify when conditions require whatever subroutine is running to be interrupted.

Simon ' s basic proposal is that emotions are " interrupt mechanisms " within the central nervous system (CNS). They have three key characteristics:

1. They interrupt ongoing goals in the CNS and substitute new goals and behaviors, including what we think of as emotional behavior.

2. They arouse the autonomic nervous system in predictable physiological ways.

3. They generate feelings of emotions.

Simon emphasizes that emotional responses are adaptive, but that does not mean that they have to be genetically determined. They can be learned, and once learned, they can be modified. Likewise, the efficacy of stimuli provoking the interrupt mechanisms can be modified through experience. As organisms learn to deal with complex situations, including social ones, they become better able to control and channel emotional responses. Some stimuli become less interruptive, others more so.

Simon ' s paper, however, is somewhat thin on detail (no specific emotion is actually discussed!) and leaves open a wide range of key questions:

- How should emotions be classified?
- Are some emotions more basic than others?
- Are there fundamental features that all emotions share?

- How exactly should we understand the role of arousal, physiological markers, and feelings in emotional experience and emotional responses?
- How far down the phylogenetic scale do emotions go?
- How can emotions be studied and measured?
- What can we say about the neural bases of emotion?

In the fifty or so years since Simon ′ s paper was published, these questions have been tackled by the growing interdisciplinary field now known as *affective science* .

Ekman and Basic Emotions

Several of the questions identified at the end of the last section received influential answers in the *basic emotion theories* , whose best-known exponent is Paul Ekman. Ekman started out from the Darwin-inspired question of whether facial expressions of emotion are cross-cultural. He realized that although there had been extensive research on the topic across different cultures, all the experimental subjects had been exposed to similar media, such as television, movies, or magazines, so the meaning of the facial expressions may have been the result of social learning rather than being psychological constants. To eliminate this possibility, Ekman conducted his groundbreaking studies in New Guinea, a preliterate culture that had been maximally visually isolated from literate cultures. His aim was to test the hypothesis that people in that culture associate emotions with facial expressions in the same way as people of literate cultures.

In the influential paper " Constants across Cultures in the Face and Emotion " (1971), Ekman and Friesen recorded the study results. Among the members of the Fore linguistic-cultural group of New Guinea, 189 adults and 130 children were recruited to participate. The participants had seen no movies, did not understand English or Pidgin (the official language in Papua New Guinea), had no contact with Westerners, and thus had minimal opportunity to learn to imitate or recognize Western facial expressions.

Subjects were shown a group of two or three photographs simultaneously, with each photo depicting a facial expression. Meanwhile, a story indicating an emotion was read (carefully chosen to reflect Fore understanding of emotions). Participants had to point to the photo in which the person ' s face showed the emotion described in the story. There were six targeted emotions: happiness, sadness, anger, surprise, disgust, and fear. They were selected because they had been found by more than one investigator to be discriminable within the literate cultures. Each subject was asked to make at least three different discriminations. To ensure consistency, the same story was told more than once, with different combinations of the correct and incorrect photos.

The results of the adult group are reflected in Table 16.1 . Each row records three emotions in a group, with the correct one in the leftmost column and the two incorrect ones in the second leftmost column.

Table 16.1 Results of facial expression recognition in New Guinea adult subjects

Adult Results

EMOTION DESCRIBED IN THE STORY	EMOTIONS SHOWN IN THE TWO INCORRECT PHOTOGRAPHS	NO. OF SUBJECTS	% CHOOSING CORRECT FACE
Happiness	Surprise, disgust	62	90**
	Surprise, sadness	57	93**
	Fear, anger	65	86**
	Disgust, anger	36	100**
Anger	Sadness, surprise	66	82**
	Disgust, surprise	31	87**
	Fear, sadness	31	87**
Sadness	Anger, fear	64	81**
	Anger, surprise	26	81**
	Anger, happiness	31	87**
	Anger, disgust	35	69**
	Disgust, surprise	35	77**
Disgust (smell story)	Sadness, surprise	65	77**
Disgust (dislike story)	Sadness, surprise	36	89**
Surprise	Fear, disgust	31	71*
	Happiness, anger	31	65*
Fear	Anger, disgust	92	64**
	Sadness, disgust	31	87**
	Anger, happiness	35	86**
	Disgust, happiness	26	85**
	Surprise, happiness	65	48
	Surprise, disgust	31	52
	Surprise, sadness	57	28[a]

* $p < .05$.

** $p < .01$.

[a] Subjects selected the surprise face (67%) at a significant level ($p < .01$, two-tailed test).

(From Table 1 in Ekman and Friesen 1971 .)

Ekman interpreted the results (for both adults and children) as clearly supporting the hypothesis that particular facial expressions are universally associated with particular emotions. In a related study, he did the reverse and asked members of literate cultures to recognize the facial expressions made by the New Guineans (see Figure 16.1). The results also support the hypothesis that there are constants across cultures in emotional facial expressions (hence the title of the paper).



Figure 16.1 Facial expressions made by the subjects from New Guinea.

(From www.ekmaninternational.com/a-brief-history-into-paul-ekmans-early-research/ .)

Ekman ' s basic emotion theory makes two main claims:

1. There are separate, discrete emotions (such as anger, fear, happiness, and so on), each of which is associated with a coherent set of responses in facial expressions, physiology, cognition, and behavior.

2. Each basic emotion serves a distinctive evolutionary function and is hardwired to occur when we deal with certain fundamental life tasks in a way that was adaptive in our evolutionary past.

Ekman ' s (most uncontroversial) list of basic emotions consists of happiness, sadness, fear, anger, disgust, and surprise. Each of these emotions organizes a coherent and distinctive set of responses. For example, when fear is activated, the facial muscles contract in a specific manner so that the eyebrows are straightly and horizontally raised, the upper eyelid is lifted, and more sclera is exposed. Presumably, these changes can help gather more information about the threat. Besides, measures of heart rate and skin conductance show that the sympathetic autonomic nervous system is highly aroused. Finally, by measuring finger temperature, a decrease in peripheral finger temperature is found. This is because blood flow is diverted away

from the periphery and redirected toward the large skeletal muscles. These changes can signal threat and prepare one to flee away from it, the paradigmatic behavioral response associated with fear. Most importantly, this set of changes makes fear distinctive from other basic emotions, as clearly shown in Figure 16.2 . As we can see, although no single physiological response can individually differentiate an emotion, the set that includes several different measures does.

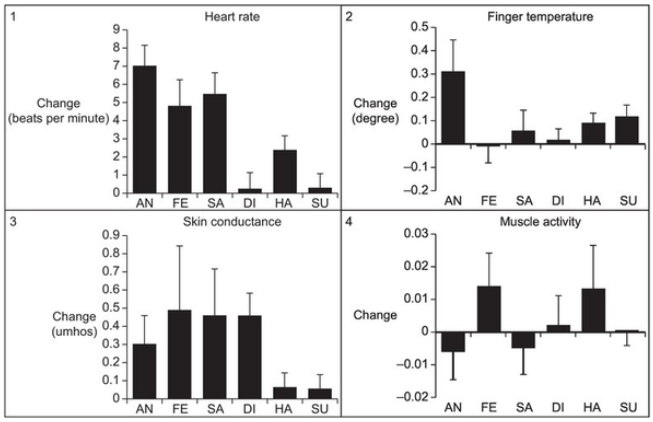


Figure 16.2 The distinct physiological responses correlated with each basic emotion. AN = anger; FE = fear; SA = sadness; DI = disgust; HA = happiness; SU = surprise.

(From Figure 3 in Levenson et al. 1990 .)

Exercise 16.2 Consider a different basic emotion (you choose!). Sketch out what a basic emotion theorist might identify as its evolutionary function and the corresponding features that help achieve this function.

Basic emotion theories have been much challenged. One powerful line of criticism comes from meta-analyses. (These are high-level studies that do

not include any new experiments but instead compare and aggregate multiple published experimental studies on a given topic.) Meta-analyses have cast doubt on the idea that there are strict links between individual emotions and specific reactions, such as facial muscles or brain neural activities. Moreover, critics from cultural anthropology have been skeptical about the idea that there are universal emotions that straddle cultures and are completely independent of social context and upbringing.

In any event, Ekman ' s work on universal facial expressions points to two important directions in emotion research. The first is studying the behavioral responses associated with particular emotions in order to shed light on their functions and possible evolutionary origins. The second is to measure brain and other physiological activity in humans and nonhuman animals to evaluate the idea that each emotion is associated with distinctive autonomic and CNS activity .

16.2 Affective Space and the Affective Scientist ' s Toolkit

Emotions are just one part of affective life. As we look in more detail at ongoing research in affective science, it is important to have a clear sense of the other components of what might be termed *affective space* . Table 16.2 distinguishes various different affective phenomena according to their typical duration and role, giving examples of each.

Table 16.2 Classifying different affective phenomena by typical duration and role

NAME	EXAMPLE	DURATION	ROLE
Emotions	Anger, joy	Short term (seconds to minutes)	Primarily operate on actions
Emotional feelings	Having a knot in one's stomach due to fear	Short term (seconds to minutes)	Bring emotions to conscious awareness
Concepts of emotions	The concept of fear	Long term	Thinking and speaking about emotions
Emotional traits	Irritability, fearlessness	Long term	Higher-order dispositional states governing emotional responses
Moods	Depression	Longer term (days to months)	Primarily operate on cognition and also indirectly on actions
Instincts	Infant crying, avoidance of painful stimuli	Short term (seconds to minutes)	Innate response patterns to specific environmental triggers
Drives	Sexual desire, hunger	Short to medium term (seconds to hours)	Associated with homeostatic disturbances

These different phenomena are not always clearly distinguished by psychologists and affective scientists. One criticism that might be leveled at Ekman ' s basic emotion theory, for example, is that his experiments are really about how people classify and categorize emotions rather than about the nature of emotions themselves.

Exercise 16.3 Go down the list of different affective states in Table 16.2 and think of a different example under each category.

Still, although there seem to be clear examples within each category, the boundaries between the categories are often fuzzy. Love would normally be classified as an emotion, although its timescale is more similar to that of moods. Many of the behaviors associated with fear seem to be instinctual. Is aggression an emotional trait, a mood, an instinct, or a drive?

Exercise 16.4 Give examples of other affective states that seem to be on the borderline between two or more different categories.

Partly because affective categories seem to blur into each other and partly because of the difficulty of drawing sharp boundaries between affective phenomena, psychologists and affective scientists typically think about affective states in general, and emotions in particular, as states that vary along two or more dimensions. Those dimensions define a space, with the location of each emotion within the space given by its coordinates (i.e., its values on each of the dimensions).

The simplest and most popular dimension model has two dimensions — valence and arousal:

Valence

This dimension is a continuum from pleasure to displeasure, measuring the attractiveness or aversiveness of a situation.

Arousal

This dimension measures the subject ' s degree of physiological/psychological engagement and responsiveness.

A well-known version of this approach is the *circumplex model of emotion* proposed by James G. Russell in 1980. A version of the model is depicted in Figure 16.3 .



Figure 16.3 The circumplex model of affect.

(From Posner et al. 2005 .)

This model was developed from factor analysis of data from self-reports, where participants were asked to categorize emotion terms according to a certain taxonomy. Note that the circumplex model includes both emotions and moods. It has sadness and depression next to each other, for example.

Increasingly, however, affective scientists are coming to the view that affective space has more than two dimensions. There are many different types of negative emotions, and it is difficult to see how all the differences between them can be explained solely in terms of their respective degrees of aversiveness and arousal. It is not hard to imagine, for example, that there could be instances of shame and anger, respectively, that were exactly the same with respect to valence and arousal.

Unsurprisingly, there are many different multidimensional models of emotions. Ralph Adolphs and David Anderson (2018), in their book *The Neuroscience of Emotion: A New Synthesis* , identify seven key features of emotional states:

Scalability . They can vary in intensity, which is similar to, but not the same as, arousal.

Valence. They vary according to pleasantness versus aversiveness.

Persistence. They outlast the stimulus that elicits them.

Generalization. They generalize over stimuli and behavior, as opposed to being fixed responses.

Global coordination . They engage the whole organism and have wide-ranging physiological/psychological expressions.

Automaticity . They can be challenging to control and regulate.

Social coordination. They serve a signaling function to conspecifics and members of other species.

Adolphs and Anderson explicitly propose this multidimensional framework as one that would allow emotions to be studied without any verbal labels. Part of their motivation is to provide a framework for emotion in both human and nonhuman animals, and so, unlike Russell ′ s circumplex model, they do not work from self-reports.

It is worth noting, though, that the multidimensional approach proposed by Adolphs and Anderson does not have much to say about the cognitive dimension of emotions (because their interests lie elsewhere, not because they don ′ t think that it exists). Anger and fear, for example, seem to represent the world in different ways. The reason that they typically engage fundamentally different patterns of behavior is that each involves interpreting the physical/social environment in very different ways — as having properties that provoke anger as opposed to properties that inspire fear .

The cognitive dimension of emotions is prominent in *appraisal theories* of emotions, first proposed in the 1960s. The basic idea of appraisal theories, as the name suggests, is that emotions involve cognitive appraisal — that is, assessing or evaluating the natural/social environment. The starting point for all versions of the theory is the thesis that emotions are adaptive responses reflecting appraisals of features in the environment relevant to the subject ' s well-being. Richard Lazarus, the most important early appraisal theorist, calls these appraisals *core relational themes* , suggesting that emotions are goal-directed mechanisms that speak to the subjects ' goals or concerns.

For appraisal theories, therefore, the dimensions of affective space have to be more directly relevant to cognitive appraisal than the dimensions proposed by Adolphs and Anderson. Table 16.3 , from a paper coauthored by appraisal theorist Jennifer Lerner, illustrates one appraisal approach.

Table 16.3 The dimensionality of emotion according to an influential version of cognitive appraisal theory

COGNITIVE APPRAISAL DIMENSIONS	ILLUSTRATIONS: NEGATIVE EMOTIONS		ILLUSTRATIONS: POSITIVE EMOTIONS	
	ANGER	FEAR	PRIDE	SURPRISE
Certainty	High	Low	High	Low
Pleasantness	Low	Low	High	High
Attentional activity	Medium	Medium	High	Medium
Anticipated effort	High	High	Low	Low
Individual control	High	Low	High	Medium
Others' responsibility	High	Medium	Low	High
Appraisal tendency	Perceive negative events as predictable, under human control, and brought about by others	Perceive negative events as unpredictable and under situational control	Perceive positive events as brought about by self	Perceive positive events as unpredictable and brought about by others
Influence on relevant outcome	**Influence on risk perception** Perceive low risk	Perceive high risk	**Influence on attribution** Perceive self as responsible	Perceive others as responsible

(From Table 1 in Lerner et al. 2015 .)

Table 16.3 clearly illustrates how this version of the appraisal theory explains the differences between the two negatively valenced emotions of anger and fear.

Exercise 16.5 Explain in your own words how the two multidimensional theories discussed in the text differ from each other.

Thinking about affective space in multidimensional terms opens up a number of promising avenues for studying emotions, as well as affective states more broadly:

- Multidimensional models allow affective scientists to study emotions independently of emotional concepts and linguistic expressions of emotion.
- The individual dimensions are scalar, so they can be measured and, in some circumstances, experimentally manipulated.
- Multidimensional models reflect the fact that affective states blend into each other and are more fine-grained than our emotional concepts might suggest.
- Multidimensional models allow for affective states that might have values of 0 on one or more dimensions. This provides tools for theorizing about emotional states in nonhuman animals, even in insects.

Elements of an Emotional Episode

To appreciate the scope and complexity of the affective scientist ′ s toolkit, it is helpful to start with a schematic representation of an emotional episode, as depicted in Figure 16.4 .

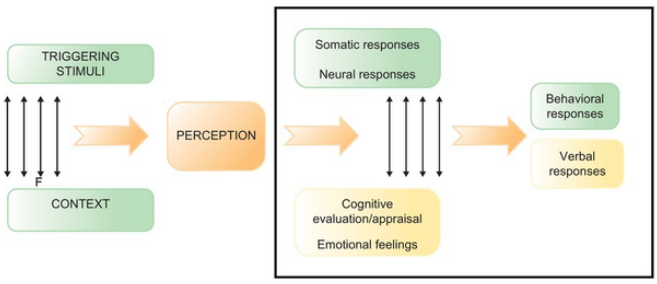


Figure 16.4 Schematic representation of an emotional episode.

Figure 16.4 is intended to do justice to the idea that emotions are widely distributed across the animal kingdom. The green boxes depict the most basic components of any emotional episode. Any emotion will normally involve a triggering stimulus that is registered perceptually and that generates a behavioral response (which might be inhibited) as a consequence of neural and somatic responses within the organism.

The yellow boxes, on the other hand, depict those components that may not be essential to every emotional episode. No nonhuman animals are capable of verbal reports. No insects are capable of cognitive evaluation or appraisal, and it seems most unlikely that they have feelings in any recognizable sense. Yet many affective scientists identify and study emotional states in insects and nonhuman animals (as we will see in more detail later in the chapter).

Affective scientists have different tools to study different components of emotional episodes, as shown in Table 16.4 .

Table 16.4 Tools for studying different components of an emotional episode

ELEMENT	TOOLS
Neural responses	Electrophysiology (single-neuron recordings) Neuroimaging (e.g., PET and fMRI) Optogenetic/pharmacogenetic interventions Genetically engineered imaging (GECIs and GEVIs) Lesion studies
Somatic responses	Standard sensors for measuring heart rate, skin conductance, body temperature, etc. Measures of facial expression (e.g., FACS)
Cognitive appraisal	Standard experimental techniques in cognitive psychology and cognitive science
Affective feelings	Psychological self-report Standardized scales and measures

FACS , Facial Action Coding System; *fMRI* , functional magnetic resonance imaging; *GECI* , genetically engineered calcium indicator; *GEVI* , genetically engineered voltage indicator; *PET* , positron emission tomography.

Some of these tools have been covered in detail in other sections of this book (e.g., electrophysiology and fMRI, as well as the experimental techniques of cognitive psychology). Some are straightforward (e.g., measuring the physiological markers of affective states). But some we have not yet encountered, and these will be discussed further in the remaining part of this chapter. We will focus in particular on the following:

- Genetic tools for studying emotions
- Using lesions to study the localization of emotions

Genetic Tools in Affective Science

Affective science is more deeply informed by genetics than other areas of cognitive science — for two reasons:

- At least since Darwin, researchers have thought that some emotions need to be understood in evolutionary terms, as evolved responses to specific types of situations. For that reason, there has been considerable research into both the genetic basis for individual

emotions and how genes can modulate the expression of emotion and account for individual differences.

- Powerful tools, such as optogenetics and pharmacogenetics, have been developed for using genetic markers to study populations of neurons. These tools can only be used in animals (primarily rodents). However, the study of emotions and affective states is one of the areas of cognitive science where animal studies are potentially most directly relevant because of commonalities in emotional response across species and the fact that many of the neural structures involved in human emotions are phylogenetically primitive.

This second reason allows us to add more tools and strategies to those surveyed in Part II.

Knockout Experiments on Emotions

Knockout experiments are a widely used tool in medical research, carried out on mice since the 1990s. The basic technique for manipulating mouse DNA is to replace a functional gene with a nonfunctional copy in mouse stem cells and then insert the modified cells into a developing embryo. This "knocks out" the relevant gene and allows researchers to investigate how the mouse's subsequent behavior might change as a consequence.

As an example, anxiety responses have been studied in knockout mice and highlighted the role of the neurotransmitter serotonin. Multiple studies have found that knocking out a particular serotonin receptor, known as *5-HT (1A)* , seems to increase anxiety levels in mice, where this is operationally defined in terms of avoidance of novel and unfamiliar stimuli (as well as

autonomic responses such as raised body temperature and increased heart rate).

Using Genes for Neuronal Interventions and Imaging

As emphasized at various points in the book, a key challenge for neuroscientists is understanding and modeling the behavior of populations of neurons. Populations of neurons are too large to be studied using the electrophysiological tools that work for individual neurons and small groups, but they are often too fine-grained to be studied with neuroimaging tools such as fMRI. Genetic tools go some way toward bridging this gap:

Optogenetics is a type of genetic engineering that allows specific groups of neurons to be controlled by light. The genetic code for the desired class of neurons is altered by adding a gene that expresses a light-sensitive ion channel (an *opsin*). These neurons can then be activated/deactivated by applying/withdrawing the appropriate type of light (one widely used opsin, channelrhodopsin-2 [ChR2], is derived from green algae and responds only to blue light). This allows neuroscientists to make targeted interventions to explore, for example, the behavioral consequences of activating specific populations of neurons.

Pharmacogenetics is a similar technique, but in this case, the engineered genes express receptors sensitive to specific synthetic drugs (and not to other neuronal activity). These receptors can be activating or inhibitory.

Optogenetics and pharmacogenetics work on different timescales. Optogenetic interventions have a resolution at the level of milliseconds,

whereas pharmacogenetic interventions can last for minutes or hours.

At the same time, genetic tools are being developed for measuring neuronal activity as a complement to electrophysiological tools. These involve engineering genes that express indirect indicators of neuronal activity, such as levels of calcium inside the cells or changes in voltage in the cell membrane. These indirect indicators, known as *genetically engineered calcium indicators* (GECIs) or *genetically engineered voltage indicators* (GEVIs), can then be measured through optical microscopes.

These types of genetic engineering have so far been used primarily in insects and rodents, although there has been some research involving optogenetics in monkeys. Nonetheless, given the importance of animal studies in understanding emotions and affective states, they are an integral part of the affective scientist ′s toolkit.

Lesion Studies

In neuroscience, lesion studies are used to investigate how damage to certain brain regions can affect specific cognitive abilities and/or behaviors. They have proved very important for understanding the biological bases of cognition.

Lesions can be studied in humans with focal brain damage, or they can be brought about experimentally in nonhuman animals. In human cases, lesions in patients can illustrate disassociations in function and thus offer valuable information about the architecture of the mind and cognition. Experimentally caused lesions can be used to explore hypotheses about the direction of information flow and how brain areas might causally influence each other.

A famous example is Phineas Gage, who had an iron rod shoot through his head in an accident in 1848. Gage did not die from the accident. However, the large lesion in the brain changed him drastically. He regained his physical capability and had no problem with sense perception or speech. What changed was his emotional behavior. To cite his doctor John Harlow ' s words, he became " fitful, irreverent, indulging at times in the grossest profanity which was not previously his custom, manifesting but little deference for his fellows, impatient of restraint or advice when it conflicts with his desires, at times pertinaciously obstinate, yet capricious and vacillating, devising many plans of future operation, which are no sooner arranged than they are abandoned " (Harlow 1868 : 277).

In the 1960s, neuroscientist Hanna Damasio used modern neuroanatomy and neuroimaging technology at the time to reconstruct Gage ' s brain. She and her colleagues inferred that the iron rod had damaged the prefrontal cortices in the ventral and inner surfaces of both the left and right hemispheres (the ventromedial prefrontal cortices). This led researchers to think that this particular area of the brain is related to emotional function. We will look at some related studies later in the chapter.

Lesion studies of human patients with brain damage are valuable in establishing the involvement of certain brain regions for emotions. However, the brain damage is usually not very specifically localized, and there is often considerable collateral damage. Lesion studies of nonhuman animals allow for much greater anatomical precision. Moreover, they also permit within-subject pre- and post-lesion comparisons. For these reasons, human lesion studies can be usefully supplemented by animal research.

There are two kinds of techniques in animal lesion studies: permanent and reversible. Scientists can create permanent lesions through aspiration or

excision, which involves performing a craniotomy to remove a portion of the skull, retracting the dura, and removing brain tissue with cautery and suction or excision. Permanent lesions can also be caused by injecting a neurotoxin, such as ibotenic acid, which can overstimulate receptors for *N* -methyl-D-aspartate (NMDA) so that glutamate and calcium are accumulated. Eventually, this leads to cell death.

Reversible techniques aim to cause temporary lesions. Pharmacological interventions involve injecting a nonneurotoxic compound into a specific brain area to temporarily reduce neurotransmission. A second technique is to apply cryogenic probes to the cortical surface, cooling the tissue to decrease neural activity. Unlike the first two techniques, the third technique is noninvasive. Transcranial magnetic stimulation (TMS) uses a magnetic inducer to the scalp. By producing electrical currents in the underlying brain tissue, brain activity can be changed.

Lesion studies can be combined with other tools. For example, the genetic tools discussed earlier can be used to create lesions. Another possibility is to combine lesion studies with fMRI data. Such data in subjects with focal lesions can show how large-scale networks change when certain nodes are damaged. This is valuable in identifying the connectivity between nodes .

16.3 Fear: A Multilevel and Multidisciplinary Case Study

The second half of this chapter explores in detail the single emotion that has probably been studied the most. This is fear, which has as good a claim as

any to be a basic emotion and is, for obvious reasons, widely distributed in the animal kingdom. Looking at fear from multiple perspectives will allow us to see the affective scientist ' s toolkit in action.

We begin with important work on fear conditioning in rodents that brings out the importance of bodily expressions of emotions and highlights the important role of the amygdala. We then turn to fear in humans, looking first at studies of patients with brain damage, then fMRI, and then neuroimaging studies.

Fear Conditioning and Fear Responses in Rodents

Conditioning experiments have long been important in cognitive science, as we saw in reviewing Tolman ' s research on place learning in Chapter 1 . One of the most studied forms of conditioning is fear conditioning. The basic experimental techniques are the same as those described in Section 1.1 .

> **Exercise 16.6** Go back to Section 1.1 and review the basic setup of conditioning experiments.

Fear conditioning is applied to what is initially a behaviorally neutral *conditioned stimulus* (CS), such as a light, a puff of air, or the sound of a tone. When the CS is paired with an aversive *unconditioned stimulus* (US), such as an electric shock to the foot, the CS comes to elicit the same behaviors as the US. As a result, the animal responds to the CS in the same way as it responds to receiving an electric shock.

The basic assumption behind fear conditioning is that the behaviors initially elicited by the US and (after training) by the CS are produced by some neural state of fear. What the animal learns is that the CS *predicts* the painful stimulus, and so the animal behaves in the way that it would if it received the US. This is a *functional* model of fear, in which fear is defined by its typical causes (the imminence of a painful stimulus) and its typical physiological and behavioral effects. Defining fear in this way allows there to be significant commonalities between states of fear in nonhuman animals and in humans.

The advantage of using fear conditioning to study fear is that it can be " switched " on easily, simply by presenting the CS, so that there is no need to place the animal in a threatening situation, which would present potential confounds and require assumptions about perception and so forth. On the other hand, of course, it rests on the assumption that the CS elicits the neural state that causes fear behavior so that the process of fear conditioning is, in essence, the same as the process by which animals acquire fear responses in " normal " situations. This assumption is widespread but rarely discussed explicitly.

Exercise 16.7 State in your own words reasons for and against using fear conditioning to study fear.

Fear conditioning studies typically employ several different physiological and behavioral markers of fear. Table 16.5 illustrates how the markers used in animal studies map onto the indicators proposed by the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*) for

diagnosing generalized anxiety (the *DSM* , published by the American Psychiatric Association and currently in its fifth version, the *DSM-5* , is a very widely used classification of mental disorders). This table illustrates both the continuities between studying emotion in humans and animals and the fluid boundaries between different types of affective disorders. Although fear and anxiety are labeled differently, they do seem to blend into each other.

Table 16.5 Measures of fear in animal studies compared to the *DSM* criteria for generalized anxiety

MEASURES OF FEAR IN ANIMAL MODELS	*DSM-III* CRITERIA—GENERALIZED ANXIETY
Increased heart rate	Heart pounding
Decreased salivation	Dry mouth
Stomach ulcers	Upset stomach
Respiration change	Increased respiration
Scanning and vigilance	Scanning and vigilance
Increased startle	Jumpiness, easy startle
Urination	Frequent urination
Defecation	Diarrhea
Grooming	Fidgeting
Freezing	Apprehensive expectation—something bad is going to happen

DSM-III , *Diagnostic and Statistical Manual of Mental Disorders* , 3rd edition.

(From Davis 1992 .)

A particular focus of fear conditioning research has been to identify the neural area(s) involved in predicting aversive stimuli. Multiple studies have pointed to the central importance of the *amygdala* in fear. The amygdala is a subcortical area that forms part of the limbic system (together with the *thalamus* , the *hypothalamus* , and the *hippocampus*).

One reason that the amygdala is important in fear and related affective states is that it has anatomical projections to multiple areas of the brain directly involved in behavioral and physiological fear responses. Some of these connections are illustrated in Figure 16.5 .

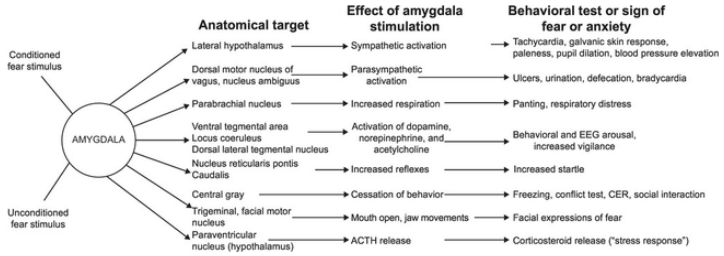


Figure 16.5 Anatomical connections from the amygdala to brain areas directly involved in characteristic behavioral and physiological fear responses. *ACTH* , adrenocorticotropic hormone; *CER* , conditioned emotional response; *EEG* , electroencephalogram.

(From Davis 1992 .)

As discussed in Chapter 9 , however, anatomical connectivity is no guarantee of either functional or effective connectivity.

Exercise 16.8 Look back at Chapter 9 and check your understanding of these three types of connectivity.

Fortunately, many experimental interventions have confirmed the role of the amygdala in fear states:

Electrical stimulation studies have shown that activity in the amygdala can produce characteristic fear responses. These include (a) autonomic responses such as increased respiration and heart rates, as well as increased blood pressure, and (b) behavioral responses, such as freezing (which is a characteristic for rodents in threatening situations).

Lesion studies show that the amygdala seems to be required for fear responses. So, for example, when rats undergo fear conditioning to a neutral stimulus and then their amygdalas are lesioned, they no longer

exhibit a freezing response to that neutral stimulus. Amygdala lesions also reduce defensive behaviors in unconditioned fear (e.g., rats with amygdala lesions will crawl over a sedated cat).

It ' s worth observing that electrical stimulation studies are evidence that the amygdala is *sufficient* to generate fear responses, whereas the lesion studies are evidence for its *necessity* .

Recent work on fear conditioning and the amygdala has been carried out on mice rather than rats (partly because such work uses the genetic tools discussed earlier, which have been primarily developed in mice). The emerging picture is much more fine-grained. For example, studies using optogenetic stimulation have shown that the basolateral nucleus (BLA), one of the areas in the amygdala most involved in fear conditioning, actually contains distinct groups of neurons that respond differentially to positive and aversive stimuli and that project to different neural regions.

In sum, there is plentiful evidence *in rodents* for the importance of the amygdala in generating autonomic and behavioral fear responses. But what does that tell us about the human case? It turns out that the amygdala is also key to fear in humans and that different nuclei in the human amygdala also function differently — but for that, we need to look at somewhat different types of evidence: studies on patients with severe amygdala damage .

Fear and Amygdala Damage in Humans

The patient who has contributed most to the modern neuroscience of emotion, especially the study of the functions of the human amygdala, is a woman called S.M. She has an extremely rare form of bilateral brain damage of the amygdala caused by Urbach – Wiethe disease (UWD), a genetic

disease that involves a mutation within the extracellular matrix protein 1 gene (*ECM1*). It is not a life-threatening disease. However, it does lead to certain functional abnormalities. S.M. has normal functioning in basic perceptual and cognitive abilities such as intelligence, memory, vision, and language. Nevertheless, she was observed to have problems in the recognition, experience, and expression of fear, as illustrated in Figure 16.6 .

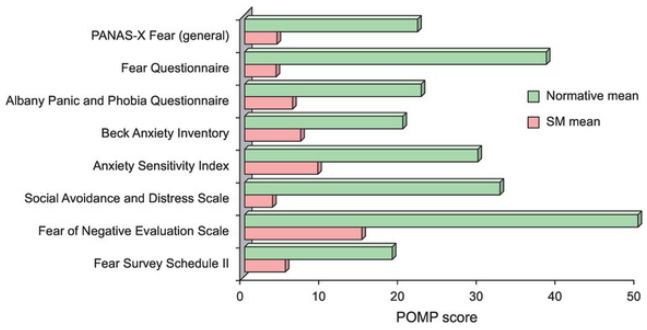


Figure 16.6 The subjective experience of fear in S.M., as measured using standard self-report tools, in comparison with normal subjects.

(From Figure 3 in Feinstein et al. 2011 .)

In a 2011 study, Justin Feinstein and colleagues exposed S.M. to live snakes and spiders (two of the most commonly feared species in the animal kingdom), took her on a tour of the Waverly Hills Sanatorium (ranked as one of the most haunted places in the world), and showed her film clips of scary movies. None of these stimuli ended up triggering the state of fear in S.M. There were no signs of avoidance behavior or withdrawal, no impairments in her conditioned skin conductance response and startle reflex, and no subjective feelings of fear based on her self-report. In fact, in all three scenarios, she insisted on poking the large and dangerous snakes and attempted to touch a tarantula, wanted to touch and tried to talk to the people dressed as monsters in the haunted hospital, and asked the name of the scary

movie so that she could watch it later herself. In related studies, S.M. appears to be capable of experiencing emotions other than fear. These results suggest that the amygdala plays a necessary role in processing fear.

As we have seen in studies with mice, the function of the amygdala is, in fact, rather fine-grained — it would be inaccurate to describe it as the "fear" area. This result in mice seems to be translatable to humans. Jack van Honk and colleagues studied a group of subjects with UWD in the Namaqualand population in South Africa. Using structural MRI to establish amygdala subregion damage and fMRI to establish remaining amygdala subregion functionality, they found five subjects with damage only in the basolateral amygdala (BLA) and not in the central and superficial nuclei. This is different from S.M., who also has damage in the central-medial amygdala (CMA). For the location of the damage in the Namaqualand patients, see Figure 16.7 .

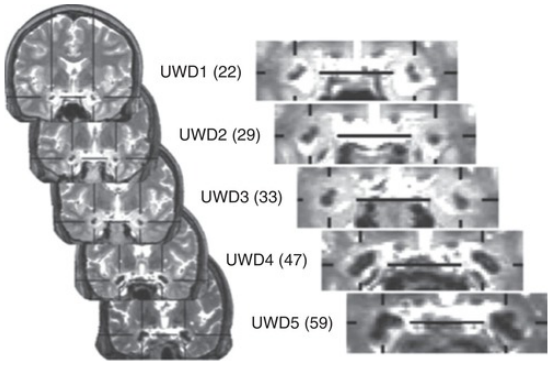


Figure 16.7 T2-weighted coronal images of the five subjects with UWD, with age in parentheses and crosshairs indicating the location of the brain damage.

(From Figure 12.2 in van Honk et al. 2016.)

Interestingly, these subjects ′ responses to fear are drastically different from those of S.M. For example, S.M. shows no fear toward common stimuli used to elicit the emotion. This can be called fear *hypovigilance* . The Namaqualand subjects with BLA damage due to UWD showed the opposite phenomenon — fear *hypervigilance* . For example, they showed hyperattention to mild, innate threat cues: fearful facial expressions, which warn the observers of potential threats of an ambiguous nature. In one such study, subjects were asked to identify several facial expressions. In comparison with the healthy control group, subjects with BLA damage with UWD performed better in the recognition of fear and looked at the fearful eyes for a longer period of time. By contrast, S.M. had difficulty paying attention to the fearful eyes and was very impaired in the recognition of fear.

This result has also been confirmed in another study that tested nonconscious fear responses through nonconscious priming.

Exercise 16.9 Review the techniques for nonconscious priming in Section 15.2 .

Subjects see nonsense images that are blue, green, or red. They are then asked to name the color as fast as possible. But they do not know that a happy, fearful, or neutral face is presented for 14 milliseconds (not consciously visible) just before the images are presented. Thus, the emotional content of the face automatically captures attention and influences the naming speed. Subjects with fear hypervigilance show attentional biases, as manifested in longer color-naming reaction times when primed with threat stimuli.

The contrast between S.M. (with damage in both the BLA and CMA) and the Namaqualand subjects (who only have BLA damage) suggests that different areas of the amygdala have specific functions. The BLA exerts regulatory control over innate fear responses, whereas the CMA acts reflexively and impulsively in fear behaviors. In the Namaqualand subjects, the damaged BLA no longer can prohibit the fear responses from being executed by the CMA. Thus, they express fear hypervigilance. Finally, in S.M., the damaged BLA and CMA cannot execute any fear responses. Thus, she appears hypovigilant. The different circuits are depicted in Figure 16.8 .

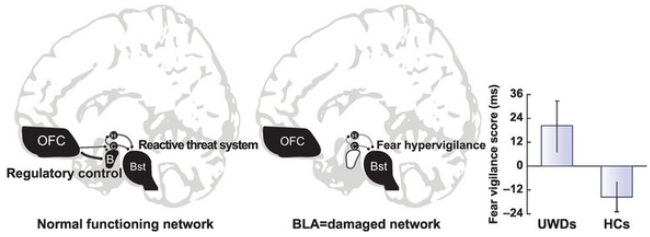


Figure 16.8 Innate threat model and fear hypervigilance in UWD. The basolateral amygdala (B) and orbito-frontal cortex (OFC) enact inhibitory control on the central-medial amygdala (C) when threat cues are mild or danger is distant. BLA damage therefore results in hypervigilant responses to innate fear cues. Bst = brain stem; H = hypothalamus.

(From Figure 12.4 in van Honk et al. 2016.)

For a better understanding of the role of the amygdala in fear processing, lesion studies can be combined with neuroimaging studies, which we look at in the next section .

Neuroimaging of Fear in Humans

Using neuroimaging to study fear in humans can be challenging. For one thing, subjects have to be in a large and noisy scanner, which places constraints on experimental design. Neuroimaging experiments typically present stimuli on a screen, but it is hard to provide ecologically valid stimuli in this way. Looking at a photograph of a facial expression flashed on a screen is very different from encountering someone in real life (where there are multiple other cues, such as body language and the situational context). Moreover, it is not always clear what exactly is being studied and measured. When subjects respond to a presented stimulus, are they actually experiencing a state of fear? Or are they classifying the stimulus in a certain way? So, for example, emotion researchers often present subjects with short stories meant to exemplify certain emotions. But the task is usually to identify the emotional state of the character in the story. This is really studying emotion concepts rather than emotions themselves.

Nonetheless, neuroimaging studies have yielded valuable insights into understanding fear (and emotions more generally). We will look at three specific strategies and representative studies from each:

1. Using a fear conditioning paradigm similar to that employed in animal studies

2. Designing experiments that are ecologically valid and can plausibly be described as eliciting an actual state of fear

3. Designing experiments explicitly focused on analyzing the *concept* of fear

Fear conditioning in humans

The first human fear conditioning experiment was actually performed just over one hundred years ago by the celebrated behaviorist John B. Watson, who conditioned a nine-month-old infant (Little Albert) to display fear behavior at the sight of a white rat, using as US the sound of a hammer hitting a pipe. This is widely held to be the first demonstration of Pavlovian conditioning in humans. Fear conditioning remains one of the most widely used paradigms for studying fear in humans. The basic technique is pretty much the same as with animals, but it uses a wider range of behavioral indices of fear, including verbal reports.

It initially proved challenging to use neuroimaging to confirm the important role of the amygdala in human fear conditioning because of the amygdala ′ s subcortical location and small size relative to the spatial resolution of fMRI (which may explain why early positron emission tomography [PET] studies failed to find increased activation in the area during fear conditioning). But those challenges have been overcome with improved technology and experimental design. So, for example, LaBar and colleagues in 1998 demonstrated that the amygdala is active in fear conditioning and extinction (when the CS is repeatedly presented without the US and the conditioned responses decline). They also found that its involvement seems to be greatest in the early stages of conditioning (presumably when the stimulus and associations are still novel).

fMRI is a particularly useful tool for exploring connectivity in the brain, and neuroimaging studies provide robust evidence for the existence of a basic network in human fear conditioning. This network, illustrated in Figure 16.9 , contains the following brain areas:

- The amygdala
- The hippocampus
- The insula
- The anterior cingulate cortex (ACC)
- The ventromedial prefrontal cortex (vmPFC)

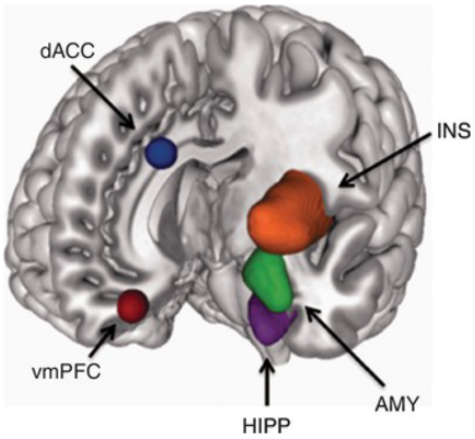


Figure 16.9 The neural regions widely held to be implicated in the human fear network.

(From Figure 1 in Maeng and Milad 2017 .)

The hippocampus is part of the limbic system and deeply involved in memory processing, whereas the insula is a part of the cerebral cortex implicated in multiple types of emotion processing (e.g., in disgust). The ACC has a dual role, with connections to both the limbic system, typically associated with emotions and homeostasis, and the prefrontal cortex, typically associated with more cognitive and executive functions. The vmPFC is a highly connected area in the prefrontal cortex thought to function as an integrative hub. In the human fear network, the ACC and vmPFC control and modulate fear responses (e.g., in extinction).

Fear conditioning has significant clinical implications, which have also been a focus of study. If fear conditioning is indeed the mechanism by which fear responses are acquired, then we might expect it to be involved in conditions that involve disordered fear responses. Posttraumatic stress disorder (PTSD) is a good example. One feature of PTSD is that fear responses overgeneralize, and thus they are easily triggered by stimuli that would not provoke them in normal subjects. Moreover, subjects with PTSD tend to show slow rates of extinction in fear conditioning paradigms. From a neurobiological perspective, studies have shown that PTSD is often characterized by hyperactivity in the amygdala and ACC, coupled with low activity in the vmPFC. Such correlations are potentially significant both for diagnosis and for evaluating treatments.

Neuroimaging fear

Further support for the existence of a human fear network comes from neuroimaging studies that elicit, in human subjects, an actual state of fear from unconditioned stimuli. This is rather difficult to do for technical and ethical reasons, but in an interesting study in 2010, Dean Mobbs and colleagues examined the neural activity associated with fear caused by a tarantula. Spiders are commonly used in fear studies because they are paradigmatic phylogenetic threats that humans under evolutionary pressures appear to have developed the innate capacity to avoid (as we saw before, spiders were also used in the study of S.M.).

Mobbs et al. pursued two main hypotheses. The first was that the increased proximity of the tarantula would progressively invoke the brain ′s fear network (controlling for whether the tarantula was approaching or

retreating). The second hypothesis was that activity in the amygdala and related areas would differentiate the tarantula ′ s approach versus retreat, irrespective of absolute distance.

To test these, they recruited twenty healthy, nonphobic volunteers. As illustrated in Figure 16.10 , these participants were asked to put one foot in a compartment, which was connected to five other compartments at varying distances. While lying in the MRI scanner bore, the participants watched a video in which a live and active tarantula was randomly placed into one of the compartments. They were actually watching prerecorded films but were unaware of that.

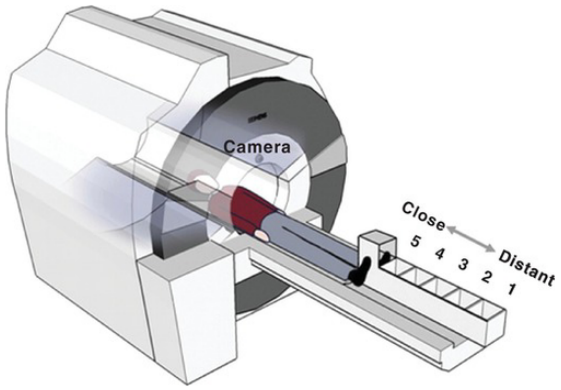


Figure 16.10 Participants in the MRI scanner bed with the compartments.

(From Figure 1 in Mobbs et al. 2010 .)

The researchers found that when the tarantula was placed closer to the participant ′ s foot, there was augmented activity in a network that included key areas discussed previously, such as the amygdala, the insula, and the ACC, as well as the *bed nucleus of the stria terminalis* (BNST; also known as the *extended amygdala*). And when the distance from the tarantula was increased, there was increased activity in the orbitomedial prefrontal cortex (omPFC) and the posterior cingulate cortex.

The results are shown in Figure 16.11 and suggest two different coping mechanisms related to fear. On the one hand, midbrain areas and the ACC correspond to an increased proximity of the threat and are thus associated with active coping such as fleeing. On the other, the omPFC is believed to be part of a pathway involved in situations with a contextually distant threat that calls for passive coping, such as freezing.

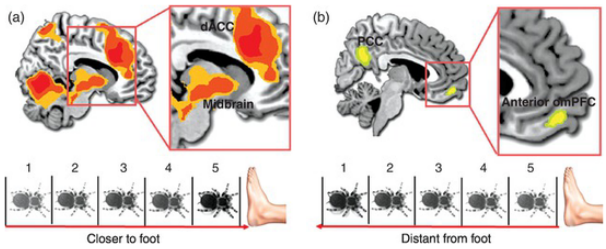


Figure 16.11 Two results from the Mobbs et al. 2010 neuroimaging study. (a) Increased proximity was associated with increased activity in the mid-brain periaqueductal gray (PAG) and dorsal ACC (dACC). (b) Increased spider distance was associated with a parametric increase in activity in the dmPFC.

(From Figure 2 in Mobbs et al. 2010 .)

They also tested the effects of altering the direction of the tarantula ' s movement. For example, for compartment 3, they compared approach to retreat by contrasting those trials in which the previous trial position was compartment 1 versus trials in which the previous position was compartment 5. Irrespective of absolute proximity, when the tarantula approached versus retreated, activity in the right amygdala, bilateral BNST, ventral striatum, and bilateral insula was observed, as shown in Figure 16.12 .

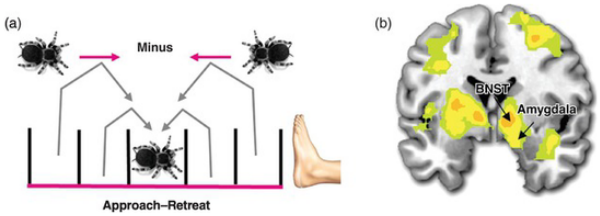


Figure 16.12 Brain activity for the retreat-versus-approach contrast in Mobbs et al. 2010 . The amygdala and BNST were differentially more active for approach versus retreat of the tarantula.

(From Figure 3 in Mobbs et al. 2010 .)

This shows that the amygdala and BNST play an important role in monitoring the trajectory of the threat and thus the oscillating value of the threat.

Neuroimaging emotion attributions

As mentioned earlier, it is not always clear in experiments on emotion whether what is being studied are emotions themselves or how people think and speak about them. So, to go back to Ekman, do his experiments provide evidence for the thesis that there are six basic emotions or, rather, for the thesis that there are six basic concepts of emotions? It might be the case, for example, that emotion concepts are discrete, whereas emotions themselves are on a multidimensional continuum. The final study that we will look at in this chapter focuses on this very question. It was carried out by Rebecca Saxe — whose work on the cognitive neuroscience of mindreading we encountered in Chapter 13 — together with colleagues at the Massachusetts Institute of Technology (MIT). Although it is not explicitly focused on fear, it nicely complements the research that we have been exploring.

As we saw earlier in this chapter, emotion researchers are moving toward thinking about the emotions in terms of a multidimensional similarity space, as opposed to a small cluster of basic emotions or a similarity space with only two dimensions (valence and arousal). Saxe and her collaborators were interested in the question of whether this is the right way to think about how we *attribute* emotions. They used neuroimaging to investigate the structure of how the brain represents emotions.

The basic idea behind their experiments was similar to that behind the experiments reviewed in Chapter 13 . They focused their attention on one part of the so-called *theory-of-mind network* .

> **Exercise 16.10** Go back to Chapter 13 and review the Saxe and Kanwisher experiments on theory-of-mind attributions.

This is the medial prefrontal cortex (MPFC), which is thought to contain non – modality-specific representations of emotions. Working with a pool of twenty different emotions identifiable through vignettes (short stories), they began by identifying patterns of neural activation in the MPFC corresponding to each of the emotional states, exploiting the fact that each emotion had ten different vignettes corresponding to it, allowing the researchers to analyze neural responses to the emotions themselves rather than to the specifics of the different stories. They then used a statistical technique called *representational similarity analysis* to construct a similarity space for the resulting activation patterns.

The experimenters used independent observers to classify the similarities between the emotions themselves according to three different

sets of dimensions. One group of observers was asked about how the emotions varied on the two dimensions of valence and arousal. A second group was given six dimensions corresponding to Ekman ' s six basic emotions, and the third group worked with thirty-eight features derived from a version of appraisal theory. The appraisal features were not themselves affective but, rather, abstract aspects of the behavior and the context (e.g., was the outcome produced intentionally?). This part of the study yielded three different similarity spaces, which were then mapped onto the neural similarity space derived from mapping activation patterns in the MPFC. It turned out that the model with thirty-eight dimensions mapped more closely onto the neural similarity space than either of the other two maps.

This study suggests, therefore, that representations and attributions of emotion may be as multidimensional and fine-grained as emotions themselves are — an intriguing note on which to end this chapter .

Summary

Affective science is a multidisciplinary and multilevel investigation of the affective aspects of mental life, which have long been underemphasized by cognitive scientists. Not least because emotions are plausibly widely distributed throughout the animal kingdom, affective scientists have a wide-ranging toolkit to work with. In addition to the standard tools for cognitive scientists, such as fMRI, affective scientists can use lesion studies in both human and nonhuman animals. Gene-editing techniques allow different types of interventions in addition to the surgical interventions that have long been used in rodents and other animals. We saw how all these tools and

techniques can be put to work by looking at the case of fear, which can be studied at multiple different levels, combining the resources of many disciplines.

Checklists

Herbert Simon developed a computational model of emotions within the general framework of the physical symbol system hypothesis.

1. The mind is a hierarchically organized information-processing system that manipulates symbols in a serial manner.

2. Any symbol system with multiple goals needs mechanisms for queuing goals and for evaluating goal completion.

3. In addition, it needs a mechanism to interrupt operational subroutines and substitute new goals and behaviors.

4. As interrupt mechanisms, emotions are adaptive but also can be learned and modified.

Ekman ' s basic emotion theory develops Darwin ' s work on facial expressions of emotion.

1. Ekman ' s New Guinea studies aimed to show that facial expressions are cross-cultural and that they correspond to a small number of basic emotions.

2. These basic emotions include happiness, sadness, anger, surprise, disgust, and fear.

3. The main claim of the basic emotion theory is that there are separate, discrete, evolutionarily functional emotions, each of which is associated with a coherent set of responses in physiology, cognition, behavior, and facial expressions.

Many affective scientists think of emotions as not having sharp boundaries but instead being located in a multidimensional space.

1. Affective space includes emotions, moods, feelings, emotional concepts, emotional traits, drives, and instincts.

2. Perhaps the most popular model of emotions (the circumplex model) has two dimensions, valence and arousal.

3. There are different multidimensional models. Adolphs and Anderson emphasize scalability, valence, persistence, generalization, global coordination, automaticity, and social coordination, whereas appraisal theorists such as Lerner emphasize the cognitive aspect of emotions.

4. Appraisal theorists ′ models rely heavily on human self-report, whereas Adolphs and Anderson are heavily influenced by animal studies.

Affective scientists have a wide range of tools at their disposal to study emotions in human and nonhuman animals.

1. Genetic tools can be used in nonhuman animals.

2. In knockout experiments on mice, a functional gene with known effects is replaced with a nonfunctional copy to investigate how behaviors change as a consequence.

3. Optogenetics involves inserting a gene expressing a light-sensitive ion channel so that researchers can activate or deactivate the neurons by controlling the light.

4. Pharmacogenetics involves inserting a gene that expresses drug sensitivity so that researchers can activate or deactivate receptors by controlling the drug.

5. Genetic tools can also be used as imaging tools to investigate indirect indicators of neuronal acidity (e.g., levels of calcium or changes in voltage in the cell membrane).

6. Lesion studies are used to investigate how damage to certain brain regions can affect specific emotional behaviors.

7. The lesion studies of Phineas Gage suggest that the ventromedial prefrontal cortices of both hemispheres are associated with emotional function.

8. Lesion studies on animals can have better specificity and control.

The most extensively used technique for studying fear is fear conditioning (a form of Pavlovian conditioning).

1. Fear conditioning pairs a behaviorally neutral conditioned stimulus (CS) with an aversive unconditioned stimulus (US) so that the subject will react the same way to both stimuli.

2. Fear conditioning in rodents reveals the importance of the amygdala.

3. Electrical stimulation of the amygdala is sufficient to produce characteristic fear responses, whereas lesion studies suggest that the

amygdala is necessary for fear responses.

4. Genetic studies in mice have shown that neurons respond differentially in different parts of the amygdala to fear stimuli.

The importance of the amygdala in fear is also shown by studies on patients, such as S.M., who have suffered amygdala damage and show impaired fear responses.

1. S.M. has bilateral brain damage of the amygdala caused by Urbach – Wiethe disease (UWD), which changes her recognition, experience, and expression of fear to the point where she appears to be *hypovigilant* (i.e., lacking responses to stimuli that typically elicit fear).

2. Studies of a group of subjects with UWD in the Namaqualand population in South Africa suggest that damage to the basolateral amygdala (BLA) area is associated with *hypervigilance* .

3. Comparing the two cases suggests that the BLA exerts regulatory control over innate fear responses, whereas the central-medial amygdala (CMA) acts reflexively and impulsively in fear responses.

Neuroimaging studies are most effective for exploring neural connectivity and the fear network.

1. Neuroimaging using functional magnetic resonance imaging (fMRI) shows that the amygdala is active in fear conditioning and extinction.

2. fMRI studies show evidence for the existence of a basic fear network that includes the amygdala, the hippocampus, the insula, the anterior

cingulate cortex (ACC), and the ventromedial prefrontal cortex (vmPFC).

3. Posttraumatic stress disorder (PTSD) is associated with hyperactivity in the amygdala and ACC and low activity in the vmPFC.

4. Neuroimaging experiments on fear of spiders show that the amygdala plays a role in monitoring the oscillating value of the threat.

5. Saxe and colleagues have shown that our representations and attributions of emotions may be better characterized by a multidimensional model than by the two-dimensional model or the basic emotion model.

Further Reading

Darwin ′ s *The Expression of the Emotions in Man and Animals* (1890) inspired many contemporary emotion researchers, particularly Ekman. Ekman 1992 is a thorough explanation and defense of his basic emotion theory. For different versions of the basic emotion theory, see the 2011 special issue of *Emotion Review* , and for criticisms, see Barrett 2006 . For Simon ′ s computational theory, see Simon 1967 . Moors and Fischer 2019 explores the related idea that emotional behaviors can be understood in terms of the same mechanisms as nonemotional behaviors.

Dukes et al. 2021 is a concise overview of the current status of affective science. Slightly longer is Fox 2018 . For more perspectives, see the papers in Fox et al. 2018 . Panksepp 2004 is an important and systematic work on the neuroscience of emotions. See also LeDoux 2012 (article) and Adolphs and Anderson ′ s book *The Neuroscience of Emotion: A New Synthesis*

(2018), which explains the logic and tools of neuro-affective science in human and nonhuman animals.

Lazarus 1991 is an important source for appraisal theories. For reviews, see Lerner et al. 2015 and Moors et al. 2013 . The latter also addresses the discrete-versus-dimensional debate, on which, see also Fontaine et al. 2007 .

Adolphs 2013 reviews the biology of fear, whereas Sehlemeyer et al. 2009 is a meta-analysis of studies showing the existence of a human fear network. Damasio et al. 1994 discusses Phineas Gage. For S.M. and other neuropsychological studies, see Bechara et al. 1994 and Bechara et al. 1999 , Feinstein et al. 2011 , and (more popular) Damasio 1994 — and for a counterweight, see Dunn et al. 2006 . Fear conditioning is reviewed in LeDoux 2000 and Shechner et al. 2014 .

For the role of the amygdala in fear, see Phelps 2006 , Feinstein et al. 2011 , and Gupta et al. 2011 . The Namaqualand studies discussed in the text come from Amaral and Adolphs 2016 , which includes other interesting papers on the amygdala in animals and humans. For the distinction between fear stimuli from outside the body and from inside, see Feinstein et al. 2013 and Gross and Canteras 2012 . The experiments on emotion concepts described in the text are in Skerry and Saxe 2015 .

Chapter 17

Looking Ahead: Challenges and Opportunities

Overview

17.1 Exploring the Connectivity of the Brain: The Human Connectome Project and Beyond

17.2 Understanding What the Brain Is Doing When It Appears Not to Be Doing Anything

17.3 Neural Prosthetics

17.4 Cognitive Science and the Law

17.5 Autonomous Vehicles: Combining Deep Learning and Intuitive Knowledge

Overview

Cognitive science has provided massive and important insights into the human mind. We have explored a good number of these in this book. These insights all stem from the single basic idea governing cognitive science as

the interdisciplinary science of the mind. This is the idea that mental operations are information-processing operations.

This book began by looking at how this way of thinking about the mind first emerged out of developments in seemingly disparate subjects, such as mathematical logic, linguistics, psychology, and information theory. Most of the significant early developments in cognitive science explored the parallel between information processing in the mind and information processing in a digital computer. As cognitive scientists and cognitive neuroscientists developed more sophisticated tools for studying and modeling the brain, the information-processing principle was extended in new directions and applied in new ways.

The interdisciplinary enterprise of cognitive science is now in excellent health. There are more contributing disciplines than ever before. Cognitive scientists have an ever-expanding range of theoretical models to work with. And there is a constant stream of technological advances in the machinery that cognitive scientists can use to study the brain. It is hard not to have a sense of optimism — a sense that cognitive science is getting close to a fundamental breakthrough in understanding cognition and the mind.

What I want to do now is to look ahead at some of the challenges and opportunities facing cognitive science at this exciting time. What follows is a small and highly personal selection of these challenges and potential applications.

17.1 Exploring the Connectivity of the Brain: The Human Connectome Project and Beyond

The successful completion of the Human Genome Project was one of the most significant scientific events of the last few decades. For the first time, scientists succeeded in identifying and mapping the 20,000 to 25,000 genes in the human gene pool, giving unprecedented insights into human genetic makeup. The Human Genome Project was so successful that it focused the minds of funding agencies on huge, collaborative projects.

In July 2009, the National Institutes of Health (NIH) announced what is, in effect, a cognitive science equivalent of the Human Genome Project: the *Human Connectome Project.* According to the funding opportunity announcement, " The overall purpose of this 5-year Human Connectome Project (HCP) is to develop and share knowledge about the structural and functional connectivity of the human brain. " Awards of $40 million were made to two consortia of research universities and institutes. The resulting collaborative and multisite effort has tackled some of the theoretical issues highlighted at various points in this book, such as the relation between different types of brain connectivity and the importance of calibrating different tools for studying the brain.

As outlined in a position paper in *Nature Neuroscience* by leading researchers in one of the two funded consortia (Glasser et al. 2016), the HCP has not only generated considerable amounts of data but also proposed a new paradigm for neuroimaging research that aims to address long-standing methodological problems with scanning and analyzing the

resulting data (some of which were identified in Chapter 9). Key elements of the paradigm proposed in Glasser et al. 2016 include the following:

- Acquiring large amounts of high-quality data on as many subjects as feasible, combining different experimental techniques
- Focusing on data with high spatial and temporal resolution and removing distortions, noise, and temporal artifacts
- Representing cortical and subcortical neuroimaging data in a common geometrical framework (termed *brainordinates*), represented in a distinctive file format (CIFTI)
- Developing a *parcellation* of the brain into distinct regions, based on connectivity and neuroanatomy
- Routinely sharing extensively analyzed results, such as statistical maps (plus raw and preprocessed data when feasible) together with the code used for the analysis, so that other neuroscientists can make precise comparisons across studies, along with replicating and extending findings

The aim and promise are to make neuroimaging more standardized and systematic. This will make comparisons across subjects and populations much easier. In particular, it will make it much easier to compare normal brains with brains that have suffered damage or disease. (For an updated atlas, see Huang et al. 2021 .)

Looking ahead, important challenges include incorporating this (or a comparable) standardizing framework across neuroimaging research and then, within that framework, developing databases on specific disorders. At the time of writing (winter 2022), ongoing projects under the auspices of

the Connectome Coordination Facility include the Alzheimer ′ s Disease Connectome Project, the Epilepsy Connectome Project, the Human Connectome Project for Early Psychosis, and Connectomes Related to Anxiety and Depression, as well as Brain Aging and Dementia.

17.2 Understanding What the Brain Is Doing When It Appears Not to Be Doing Anything

One of the themes of the Human Connectome Project has been the importance of studying the brain ′ s resting state. This is another frontier for neuroscience.

Neuroimaging and electrophysiological experiments standardly explore what happens in the brain when certain very specific tasks are being carried out. So, for example, neuroimaging experiments typically identify the different brain areas where the blood-oxygen-level – dependent (BOLD) contrast is highest during a given task. This is the basis for inferences about the localization of function in the brain. But, some researchers have argued, task-specific activation is simply the tip of the iceberg. Marcus Raichle and colleagues at Washington University in St. Louis have argued that we shouldn ′ t just pay attention to departures from the baseline set by the brain ′ s default mode of operation. There is a huge amount of activity going on in the brain even when subjects are resting with their eyes closed or passively looking at a fixed stimulus. This default mode of brain function has not yet been systematically studied by neuroscientists, but it might be quite fundamental to understanding cognition. Concentrating solely on task-

dependent changes in the BOLD signal may turn out to be like trying to understand how tides work by looking at the shape of waves breaking on the shore.

What is now often called the *default-mode network* (DMN) can be studied in pure resting-state experiments, where subjects are imaged while not performing any directed task. The brain areas most frequently identified in such experiments include the medial posterior cortex, particularly the posterior cingulate cortex and the precuneus, and the medial frontal cortex, in addition to areas around the temporoparietal junction (TPJ) area.

One interesting possibility starting to gain traction is that some cognitive disorders and diseases may be correlated with impaired functioning of the DMN. Numerous studies have observed significant correlations between deteriorating connectivity of the DMN over time and two well-known markers of early Alzheimer ′ s: rising levels of amyloid beta (the key component of brain plaques in Alzheimer ′ s) and pathologies of tau protein (which forms tangles inside neurons and disturbs synaptic communication). For large studies see Chhatwal et al. 2013 , Thomas et al. 2014 , and Buckley et al. 2017 . Epilepsy, Parkinson ′ s, and attention deficit − hyperactivity disorder (ADHD) are other disorders where impaired functioning of the DMN may be important, as reviewed in Mohan et al. 2016 . Chow et al. 2021 explores the connectivity of the DMN in healthy aging, whereas Veselinovi ć et al. 2021 considers more general connections between resting-state connectivity and cognition.

17.3 Neural Prosthetics

Suppose that, as many cognitive scientists think, important cognitive functions are carried out by functionally specialized systems that are themselves implemented in specific neural locations. Wouldn ′ t it then be possible to build mechanical devices that could replace a damaged system in the brain, reversing the effects of disease or injury? Cognitive science certainly predicts that this type of *neuroprosthesis* ought to be possible. If cognitive systems are computational devices, whose job is basically transforming a certain type of input into a certain type of output, then the crucial thing to work out is how the input and output are represented in the brain and what the basic transformations are. If this can be done, then the only obstacles to building neuroprostheses are technological.

In fact, some types of neuroprostheses are already widely used. Cochlear implants can restore hearing to individuals with hearing problems — even to the profoundly deaf. They work by providing direct electrical stimulation to the auditory nerve (doing the job that would otherwise be done by hair cells in the cochlea, which is in the inner ear). This is an input prosthetic.

Output prosthetics are much more complicated. Scientists and engineers in Theodore Berger ′ s lab at the University of Southern California have been working on a pioneering example of an output prosthetic. They have been designing and building a prosthetic implant to restore normal functioning when the hippocampus is damaged (the hippocampus plays an important role in forming and storing memories). The aim is to develop a device that will measure electrical inputs to the hippocampus, calculate what outputs would typically be generated in

normal subjects, and then stimulate areas of the hippocampus to mimic a normally functioning brain. When the second edition of this book was being written (August 2013), an early prototype hippocampal prosthetic had been tested in rats (Berger et al. 2012) and monkeys. Now, nine years later, the first tests have been carried out on humans, showing a significant improvement in episodic memory in patients with epilepsy (Hampson et al. 2018).

Exoskeletons (robot suits) are an even more dramatic example of output prosthetics. Neuroscientists, working together with biomechanical engineers, have produced motor prostheses that restore some movement to paralyzed patients. At the opening ceremony of the 2014 Soccer World Cup in Brazil, a young Brazilian paraplegic wearing an exoskeleton got out of his wheelchair and kicked the ball to demonstrate the rehabilitation possibilities. The exoskeleton was designed by a team led by Gordon Cheng at the Technical University in Munich. The hydraulically powered suit is built from lightweight alloys. The user ' s electroencephalogram (EEG) waves are read and translated into signals that control the robot ' s limbs. At the same time, feedback is sent to the user ' s brain from sensors in the soles of the robot ' s feet. The brain – computer interface making this all possible was developed by Miguel Nicolelis at the Duke University Medical Center.

For an open-access review of brain – computer interfaces in rehabilitative medicine, see Lazarou et al. 2018 . Many brain – computer interfaces use EEG technology (for reviews, see Lebedev et al. 2018 and Abiri et al. 2019).

17.4 Cognitive Science and the Law

The intensely interdisciplinary nature of cognitive science has been a recurring theme in this book. We have looked at how cognitive science has been molded by contributions from psychology, philosophy, neuroscience, linguistics, computer science, and mathematics — to give just a partial list. But the list of disciplines for which cognitive science is potentially relevant is even longer. One area where the dialog with cognitive science is gaining momentum is the law.

There are many points of contact between cognitive science and the law. Eyewitness testimony is a good example. Eyewitness testimony is a fundamental pillar of almost every legal system, including retrospective identification through line-ups and similar techniques. Yet there is strong evidence that eyewitness testimony is both unreliable and manipulable (see Busey and Loftus 2007), leading to serious miscarriages of justice (subsequently discovered, for example, through DNA evidence). For those and other reasons, the National Academy of Sciences convened an expert panel to study how identification errors arise in eyewitness testimony.

The resulting report, *Identifying the Culprit: Assessing Eyewitness Identification* , released in 2014, made a number of recommendations for courts and for law enforcement agencies. For example:

1. Line-ups should be administered by " blinded " managers who have no knowledge of which participant is suspected of the crime.

2. Witness instructions should be standardized and designed to yield a consistent and conservative response.

These and other recommendations were incorporated into US Justice Department guidelines in 2017. See Albright 2017 for a review of the scientific issues emerging from the report and some of its specific recommendations.

Neuroscientific studies and findings are increasingly being used in court cases, and this has given rise to the area known as *neurolaw* , where cognitive science intersects with forensic psychiatry, as well as legal practice. For a range of perspectives on neurolaw, see the essays in Morse and Roskies 2013 .

Some examples of questions actively being explored in this area are as follows:

- How can courts use neuroscientific findings to adjudicate questions about competence and capability that may arise, for example, in end-of-life issues?
- Are there prospects for developing brain-based techniques of lie detection?
- How should neuroscientific evidence be used as part of an insanity defense or as evidence of mitigating circumstances?
- Does the use of neuroscientific evidence in courtrooms pose potential challenges to privacy rights?
- Can neuroscience be used to predict recidivism in offenders?
- Are there possibilities for neuroscience-based interventions that will diminish the risk of an offender reoffending (e.g., deep-brain stimulation as a way of reducing sexual drive in sexual offenders)?

Some of these issues remain in the realm of theory. But others have had serious practical ramifications. In *Miller v. Alabama* , decided in 2012, the US Supreme Court decided that mandatory life sentences without the possibility of parole are unconstitutional for juvenile offenders. Justice Elena Kagan, writing for the majority, cited an amicus brief from the American Psychological Association summarizing a range of evidence from neuroscience and developmental psychology on adolescent brain development and associated vulnerabilities. In *Montgomery v. Louisiana* , decided in 2016, the High Court applied its earlier decision retroactively, potentially affecting more than 2,000 currently incarcerated individuals.

For recent research on biomarkers of criminal behavior, see Darby et al. 2018 , Savopolous and Lindell 2018 , and Hofhansel et al. 2020 .

17.5 Autonomous Vehicles: Combining Deep Learning and Intuitive Knowledge

Research on self-driving cars has been one of the most visible products of the deep-learning revolution in artificial intelligence (AI). Self-driving cars have been tested on semipublic roads, and both large corporations and small start-ups are making ambitious predictions about when they will be widely available. In fact, 2018 was the target date proposed in 2015 both by Elon Musk of Tesla and by Google. But a series of widely publicized crashes, some fatal, have put somewhat of a dampener on the enthusiasm.

As we saw in Chapter 12 , deep-learning algorithms have exponentially increased the power of machine learning, both with and without supervision. For many cognitive scientists and other observers,

though, deep learning has its limitations. The great successes of deep learning have all been in relatively circumscribed domains. Chess and Go are obvious examples. But so, too, is image recognition. Image recognition is a purely passive activity. It is a matter of identifying patterns in a data set and then projecting those patterns onto new exemplars. Admittedly, pattern recognition is no mean achievement. Traditional AI and machine learning were only able to make limited progress for many years, whereas deep-learning algorithms can outperform human experts on many complex tasks.

But still, many have thought that self-driving cars (and other forms of autonomous vehicles, such as submarines, planes, and drones) need more than sensitivity to patterns and the ability to learn from experience. They need to be able to deal with the unexpected — completely unpredictable behavior from other drivers, pedestrians, and cyclists (not to mention wild animals and livestock). Will it always be possible for a self-driving car to deal with unpredictable situations by extrapolating from its training set? Boosters for current designs for self-driving cars think so, but others are not so sure.

Some cognitive scientists see a need for self-driving cars to have an analog of the intuitive knowledge that humans call *common sense* . Specifically, they observe that human drivers are constantly exploiting their knowledge of physical objects and how objects move and behave, as well as their knowledge of other drivers and road-users. In other words, human driving exploits some of the key abilities that we have looked at in earlier chapters — what we called *folk physics* in Chapter 11 and *mindreading* in Chapter 13 . So, a key challenge, perhaps *the* key challenge, for designers of self-driving cars is how to equip their vehicles with this kind of very general knowledge.

It may turn out that folk physics and mindreading abilities ultimately rest on very sophisticated forms of pattern recognition, and thus learning them would be within the reach of some version of deep-learning algorithms (although perhaps ones that differ significantly from existing algorithms and networks). Alternatively, they may depend on fundamentally different forms of learning. This second possibility is where a spin-off company of the Massachusetts Institute of Technology (MIT) called iSee is placing its bets. The company is developing forms of probabilistic programming inspired by Bayesian models of the mind to equip self-driving cars with the common sense that they currently seem to lack . [1]

These are just some of the exciting challenges and opportunities opening up for cognitive scientists in the years ahead. I hope that readers of this book will pursue these — and, I hope, develop others.

[1] The founders of iSee are associated with Joshua Tenenbaum ' s lab at MIT. We looked at his work in Chapter 12 . There is a short article on iSee in the online *MIT Technology Review* (published on September 20, 2017).

Glossary

abduction (abductive reasoning):
a form of reasoning in which one derives a conclusion as the best explanation of given evidence, even though it is not entailed by the evidence that it explains.

absolute judgment:
a judgment about the intrinsic properties of a stimulus (e.g., naming a color or identifying the pitch of a particular tone), as opposed to a relative judgment comparing two stimuli.

access consciousness (or A-consciousness):
information available or " poised " for conscious thought and action.

action potentials:
electrical impulses fired by neurons down their axons to other neurons.

activation function:
a function that assigns an output signal to a neural network unit on the basis of the total input to that unit.

algorithm:

a finite set of unambiguous rules that can be systematically applied to an object or set of objects to transform it or them in definite ways in a finite amount of time.

amygdala:
a subcortical area in the limbic system that is usually associated with emotion processes, particularly fear.

anatomical connectivity:
the anatomical connections between different brain regions.

anterograde amnesia:
the loss of memory of events after the onset of a brain injury.

appraisal:
in theories of emotion, this is the assessment or evaluation of the natural or social environment in terms of the subject ′s well-being.

arousal:
the dimension that characterizes an emotion by measuring the subject ′s degree of physiological and psychological engagement and responsiveness.

artificial neural network (connectionist network):
an abstract mathematical tool for modeling cognitive processes that uses parallel processing between intrinsically similar units (artificial neurons) organized in a single- or multilayer form.

attractor:

a region in the state space of dynamical systems on which many different trajectories converge.

autoencoder:
type of unsupervised neural network important in deep learning. It can learn to represent features by compressing data through a bottleneck.

backpropagation algorithm:
a learning algorithm in multilayer neural networks in which error is spread backward through the network from the output units to the hidden units, allowing the network to modify the weights of the units in the hidden layers.

basic emotions:
discrete emotions (such as anger, sadness, etc.), each of which is claimed to be associated with a coherent set of responses in facial expressions, physiology, cognition, and behavior, as well as to serve a distinctive evolutionary function.

Bayes ' s rule:
gives a way of calculating the posterior probability of a hypothesis, conditional on some evidence. To apply it, you need to know the prior probability of the hypothesis and the likelihood of the evidence, conditional on the hypothesis.

Bayesianism:
movement in statistics that interprets probability subjectively and holds that rational thinkers will update their probabilities according to Bayes ' s

rule and that rational agents will maximize expected utility.

behavior-based robotics:
movement in robot design that moves beyond purely reactive subsumption architectures by allowing robots to represent their environment and plan ahead.

behaviorism:
the school of psychology holding that psychologists should only study observable phenomena and measurable behavior. Behaviorists maintain that all learning is the result of either classical/Pavlovian or operant conditioning.

binding problem:
the problem of explaining how information processed in separate neural areas of the information-processing pathway is combined to form representations of objects.

binocular rivalry:
phenomenon that occurs when different images are presented to each eye. Subjects experience an alternation of the images.

biorobotics:
the enterprise of designing and building models of biological organisms that reflect the basic design principles of those organisms.

bit:

a measure of the information necessary to decide between two equally likely alternatives. For decisions between *n* alternatives, the number of bits = $\log 2^{n}$.

blindsight:
a neurological disorder typically resulting from lesions in the primary visual cortex (V1, or the striate cortex). Like patients with unilateral spatial neglect, patients with blindsight report little to no awareness in one side of their visual field.

BOLD signal:
the blood-oxygen-level – dependent (BOLD) signal measures the contrast between oxygenated and deoxygenated hemoglobin in the brain, generally held to be an index of cognitive activity. The increase in blood oxygen can be detected by a functional magnetic resonance imaging (fMRI) scanner because oxygenated and deoxygenated hemoglobin have different magnetic properties.

Boolean function:
a function that takes sets of truth values as input and produces a single truth value as output.

Brodmann areas:
different regions of the cerebral cortex identified by neurologist Korbinian Brodmann. The primary visual cortex, for example, is Brodmann area 17.

cerebral cortex:

the parts of the brain, popularly called " gray matter, " that evolved most recently.

channel capacity:
the maximum amount of data that an information channel can reliably transmit.

chatbot:
a program set to respond to certain cues by making one of a small set of preprogrammed responses; these programs cannot use language to report on or navigate their environments because they do not analyze the syntactic structure or meaning of the sentences they encounter.

cheater-detection module:
hypothetical cognitive system specialized for identifying a " free rider " in a social exchange (i.e., a person who is reaping benefits without paying the associated costs).

chunking:
Miller ' s method of relabeling a sequence of information to increase the amount of data that the mind can reliably transmit, for example relabeling sequences of digits with single numbers (i.e., 1100100 becomes " 100 ").

Church – Turing thesis:
the thesis that the algorithmically calculable functions are exactly the functions that can be computed by a Turing machine.

circumplex model of emotion:
two-dimensional model of emotion proposed by James Russell. The dimensions are valence and arousal.

---

classical/Pavlovian conditioning:
the process of creating an association between a reflex response and an initially neutral stimulus by pairing the neutral stimulus (e.g., a bell) with a stimulus (e.g., food) that naturally elicits the response (e.g., salivation).

---

competitive network:
an example of an artificial neural network that works by unsupervised learning.

---

computation:
purely mechanical procedure for manipulating information.

---

computational neuroscience:
the use of abstract mathematical models to study how the collective activities of a population of neurons could solve complex information-processing tasks.

---

conditional probability:
the probability that some proposition A is true, on the assumption that some other proposition is true, for example the probability that it is raining (A) if we assume that it is cloudy (B). Written as p (A/B).

---

congruence priming:

a priming task in which the basic category of a prime (e.g., a tool) enhances the salience of other stimuli matching that category (e.g., other tools).

connectionist network:
see *artificial neural network* .

connectivity, anatomical:
physiological connections between segregated and distinct cortical regions.

contralateral organization:
occurs when each hemisphere of the brain processes input information from the opposite side of space (e.g., when an auditory stimulus presented to the right ear is processed by the left hemisphere of the brain).

convolutional neural network:
type of deep-learning neural network that is very important in machine vision. Characterized by sparse connectivity, shared weights, and invariance under translation. Particularly well suited to image recognition.

co-opted system:
according to simulation theory, a system specialized for a specific cognitive task that is then used to perform related mindreading tasks.

corpus callosum:

the large bundle of fibers connecting the two hemispheres of the brain.

counterfactual:
a statement about what would have happened had things been different.

covert attention:
the possibility of directing attention at different peripheral areas while gaze is fixated on a central point.

cross-lesion disconnection experiments:
experiments designed to trace connections between cortical areas in order to determine the pathways along which information flows. These experiments take advantage of the fact that the brain is divided into two hemispheres, with the major cortical areas being the same on each side.

cross-talk:
the process in which separate subsystems collaborate in solving information-processing problems using each other ′s outputs as inputs.

decision trees:
a branching representation of all possible paths through a problem space starting from an initial point.

deep learning:
very impactful approach to machine learning using multilayered and hierarchically organized artificial neural networks to achieve representation learning.

deep structure:

in Chomskyan linguistics, the deep structure of a sentence is its " real " syntactic structure, which serves as the basis for fixing its meaning. Two sentences with different surface structures can have the same deep structure (e.g., " John kissed Mary " and " Mary was kissed by John ").

dichotic listening experiments:
experiments in which subjects are presented with information in each ear in order to investigate selective attention in the auditory system.

dishabituation paradigm:
a method for studying infant cognition that exploits the fact that infants look longer at events that they find surprising.

distributed representation:
occurs when (as in many connectionist networks) objects or properties are represented through patterns of activation across populations of neurons rather than through individual and discrete symbols.

domain specific:
term used to characterize cognitive mechanisms (modules) that carry out a very specific information-processing task with a fixed field of application.

dorsal pathway:
the neural pathway believed to be specialized for visual information relevant to locating objects in space. This pathway runs from the primary visual cortex to the posterior parietal lobe.

double dissociation:
experimental discovery that each of two cognitive functions can be performed independently of the other.

dynamical systems hypothesis:
radical proposal to replace information-processing models in cognitive science with models based on the mathematical tools of dynamical systems theory.

dynamical systems theory:
branch of applied mathematics using difference or differential equations to describe the evolution of physical systems over time.

early-selection model:
a cognitive model of attention in which attention operates as a filter early in the perceptual process and acts on low-level physical properties of the stimulus.

effective connectivity:
the causal flow of information between different brain regions.

electroencephalography (EEG):
experimental technique for studying the electrical activity of the brain.

entropy:
a measure of how well a particular attribute classifies a set of examples. The closer the entropy is to 0, the better the attribute classifies the set.

event-related fMRI:
neuroimaging technique for measuring the blood-oxygen-level – dependent (BOLD) signal associated with rapidly changing neural events, which is possible because of the linear nature of the hemodynamic response.

event-related potentials (ERPs)/event-related magnetic fields:
cortical signals that reflect neural network activity that can be recorded noninvasively using electroencephalography (EEG) or magnetoencephalography (MEG).

expected utility:
the expected utility of an action is the sum of the utility anticipated from each of its possible outcomes, each discounted by its probability. In Bayesianism, rational agents maximize expected utility.

expert systems research:
a field of artificial intelligence that aims to reproduce the performance of human experts in a particular domain.

false-belief task:
an experimental paradigm first developed by psychologists Heinz Wimmer and Joseph Perner, exploring whether young children understand that someone might have mistaken beliefs about the world.

feature engineering:
the process in machine learning of classifying a database in terms of

relevant features.

feature learning:
see *representation learning.*

feedforward network:
a connectionist network in which activation spreads forward through the network; there is no spread of activation between units in a given layer or backward from one layer to the previous layer.

fixed neural architectures:
the identification of determinate regions of the brain associated with particular types of modular processing.

folk physics:
an intuitive understanding of some of the basic principles governing how physical objects behave and interact.

formal property:
a physical property of a representation that is not semantic (e.g., a formal property of the word *apple* is that it is composed of five letters of the English alphabet).

fovea:
area in the center of the retina where visual acuity is highest, corresponding to the center of the visual field.

frame problem:

the problem of developing expert systems in artificial intelligence and building robots that can build into a system rules that will correctly identify what information and which inferences are relevant in a given situation.

functional connectivity:
the statistical dependencies and correlations between activation in different brain areas.

functional decomposition:
the process of explaining a cognitive capacity by breaking it down into subcapacities that can be analyzed separately. Each of these subcapacities can in turn be broken down into further nested subcapacities, until the process bottoms out in noncognitive components.

functional magnetic resonance imaging (fMRI):
technology for functional neuroimaging that measures levels of blood oxygen as an index of cognitive activity.

functional neuroimaging:
a tool that allows brain activity to be studied noninvasively while subjects are actually performing experimental tasks (e.g., positron emission tomography [PET], functional magnetic resonance imaging [fMRI]).

functional system:
a system that can be studied and understood primarily in terms of the role it plays and the task that it executes, irrespective of the mechanism of

implementation. These systems are studied only at the computational level and are multiply realizable. (See also *multiple realizability* .)

global workspace theory of consciousness:
a leading theory of how mental states become conscious. According to this theory, attention makes low-level modular information available to conscious control (the " global workspace "), where the information is then " broadcast " to other areas of the brain.

GOFAI:
good old-fashioned artificial intelligence — as contrasted, for example, with artificial neural networks or biorobotics.

graceful degradation:
the incremental deterioration of cognitive abilities that is imperceptible within small time frames.

halting problem:
the problem first raised by David Hilbert of algorithmically determining whether a computer program will halt (i.e., deliver an output) for a given input.

hard problem of consciousness:
the problem of explaining phenomenal consciousness by appealing to physical processes in the brain and using the traditional tools of cognitive science.

Hebbian learning:

Donald Hebb′s model of associative processes, according to which "neurons that fire together, wire together."

heuristic search hypothesis:
Newell and Simon′s hypothesis that problems are solved by generating and algorithmically transforming symbol structures until a suitable solution structure is reached.

hidden layer:
a layer of hidden units in an artificial neural network.

hidden unit:
a unit (artificial neuron) in an artificial neural network whose inputs come from other units and whose outputs go to other units.

information channel:
a medium that transmits information from a sender to a receiver (e.g., a telephone cable or a neuron).

informational encapsulation:
property of modular systems that operate with a proprietary database of information and are insulated from background knowledge and expectations.

integration, principle of:
fundamental idea of neuroscience stating that cognitive function involves the coordinated activity of networks of different brain areas, with different types of tasks recruiting different types of brain areas.

intentional realism:
the thesis that propositional attitudes (e.g., beliefs and desires) can cause behavior.

intentionality:
property by virtue of which symbols represent objects and properties in the world.

interocular suppression:
a technique used to study consciousness, in which one eye is presented with an image of an object while the other eye is presented simultaneously with a high-contrast pattern that blocks conscious awareness of the presented image.

interrupt mechanism:
Herbert Simon claims that emotions function as interrupt mechanisms monitoring the environment to identify when conditions require whatever subroutine is running to be interrupted.

joint visual attention:
occurs when infants look at objects and take pleasure in doing so because they see that another person is both looking at that object and noticing that the infant is also looking at the object.

knockout experiments:
experiments that use a genetic technique that replaces a functional gene with a nonfunctional copy in cells.

Knowledge Argument:
a thought experiment proposed by Frank Jackson and featuring a neuroscientist called Mary who is confined to a black-and-white room and has never experienced colors. Mary knows all the physical facts there are to be known, and yet, according to Jackson, there is a fact that she discovers when she leaves the room — the fact about what it is like for someone to see red.

---

language-of-thought hypothesis:
a model of information processing developed by Jerry Fodor that holds that the basic symbol structures that carry information are sentences in an internal language of thought (sometimes called *Mentalese*) and that information processing works by transforming those sentences in the language of thought.

---

late-selection model:
a cognitive model of attention in which attention operates as a filter on representations of objects after basic perceptual processing is complete.

---

lesion studies:
studies in neuroscience that investigate how damage to certain brain regions can affect specific cognitive abilities and/or behaviors.

---

lexical access:
the processing involved in understanding single words.

---

likelihood:

important concept for Bayes ′ s rule, measuring the probability of some evidence E, conditional on a hypothesis H — that is, how likely you think it is that you would see the evidence if the hypothesis were true. It is standardly written as the conditional probability *p* (E/H).

linear separability:
characteristic of Boolean functions that can be learned by neural networks using the perceptron-convergence learning rule.

local algorithm:
a learning algorithm in a connectionist network in which an individual unit weight changes directly as a function of the inputs to and outputs from that unit (e.g., the Hebbian learning rule).

local field potential (LFP):
an electrophysiological signal believed to be correlated with the sum of inputs to neurons in a particular area.

locus-of-selection problem:
the problem of determining whether attention is an early-selection phenomenon or a late-selection phenomenon.

logical consequence:
a conclusion is the logical consequence of a set of premises only if there is no way of interpreting the premises and conclusion that makes the premises all true and the conclusion false.

logical deducibility:

one formula is logically deducible from another only if there is a sequence of legitimate, formal steps that lead from the second to the first.

machine learning:
the production of an algorithm that will organize a complex database in terms of some target attribute by transforming symbol structures until a solution structure, or decision tree, that will clarify incoming data is reached.

machine learning algorithm:
an algorithm for constructing a decision tree from a vast database of information.

magnetoencephalography (MEG):
brain-imaging technique that measures electrical activity in the brain with magnetic fields.

mandatory application:
a feature of modules where cognitive modules respond automatically to stimuli of the appropriate kind. They are not under any level of executive control.

masked priming:
a priming task in which a stimulus is made invisible by presenting a second stimulus (the mask) in rapid succession.

massive modularity hypothesis:

holds that all information processing is carried out by specialized modules that emerged in response to specific evolutionary problems (e.g., cheater-detection module).

mental architecture:
a model of the mind as an information processor that answers the following three questions: In what format is information carried in a cognitive system? How is information in the cognitive system transformed? How is the mind organized to function as an information processor?

metarepresentation:
metarepresentation occurs when a representation is used to represent another representation rather than to represent the world (e.g., a representation of another person ′s mental state).

micro-world:
an artificially restrictive domain used in artificial intelligence in which all objects, properties, and events are defined in advance.

mirror neurons:
neurons in monkeys that fire both when the monkey performs a specific action and when it observes that action being performed by an observer.

module:
cognitive system dedicated to performing a domain-specific information-processing task. Typically held to be informationally encapsulated but not necessarily to have a fixed neural architecture.

morphological computation:
a research program in robotics for minimizing the amount of computational control required in a robot by building as much as possible of the computation directly into its physical structure.

multilayer network:
an artificial neural network containing one or more hidden layers.

multiple realizability:
a characteristic of functional systems whose tasks can be performed by a number of different physical manifestations. For example, a heart, when viewed as a functional system, is multiply realizable because human hearts and mechanical hearts can perform the same function.

neuroeconomics:
interdisciplinary area where concepts and tools from economics are used to illuminate brain functioning.

neurotransmitters:
neurochemicals that are transmitted across synapses in order to relay, amplify, and modulate signals between a neuron and another cell.

object permanence:
the knowledge that an object exists even when it is not being perceived — an important milestone in children ' s development.

operant conditioning:

a type of conditioning in which an action (e.g., pushing a lever) is reinforced by a reward (e.g., food).

optogenetics:
a type of genetic engineering that allows specific groups of neurons to be controlled by light.

overregularization errors:
systematic mistakes that children make during the process of language acquisition as they begin to internalize basic grammar rules. Children apply rules (such as adding the suffix " -s " to nouns to make them plural) to words that behave irregularly (e.g., saying " foots " instead of " feet ").

paired-image subtraction paradigm:
an experimental technique that allows neuroimagers to identify the brain activation relevant to a particular task by filtering out activation associated with other tasks.

parallel processing:
simultaneous activation of units in an artificial neural network that causes a spread of activation through the layers of the network.

perceptron:
a single-unit (or single-layer) artificial neural network.

perceptron-convergence rule (delta rule):

a learning algorithm for perceptrons (single-unit networks). It changes a perceptron ′s threshold and weights as a function of the difference between the unit ′s actual and intended output.

pharmacogenetics:
a type of genetic engineering that allows specific groups of neurons to be controlled by synthetic drugs.

phenomenal consciousness (or P-consciousness):
the experiential or " what it ′s like " aspect of consciousness (e.g., the distinctive experience of smelling a rose or touching a piece of velvet cloth).

phrase structure grammar:
describes the syntactic structure of a natural language sentence in terms of categories such as verb phrase and noun phrase. Permissible combinations of syntactic categories are given by phrase structure rules, for example the rule stating that every sentence must contain both a verb phrase and a noun phrase.

physical symbol system:
a set of symbols (physical patterns) that can be combined to form complex symbol structures and contains processes for manipulating symbol structures. These processes can themselves be represented by symbols and symbol structures within the system.

physical symbol system hypothesis:

Newell and Simon ′s hypothesis that a physical symbol system has the necessary and sufficient means for general intelligent action.

positron emission tomography (PET):
a functional neuroimaging technique in which localization of cognitive activity is identified by measuring blood flow to specific areas of the brain.

posterior probability:
this is the result of applying Bayes ′s rule. It is the probability of a hypothesis H, conditional on some evidence E (i.e., the conditional probability p[H/E]).

poverty-of-stimulus argument:
maintains that certain types of knowledge must be innate because they are too complicated to be learned from the impoverished stimuli to which humans are exposed (e.g., Chomsky ′s argument for universal grammar).

pragmatics:
the branch of linguistics concerned with the practical implication of language and what is actually communicated in a given context.

predicate calculus:
formal system for exploring the logical relations between formulas built up from symbols representing individuals, properties, and logical operations. Unlike the propositional calculus, the predicate calculus includes quantifiers (ALL or SOME) that allow representations of generality.

prestriate cortex:
an area in the occipital and parietal lobes that receives cortical output from the primary visual cortex.

primary visual cortex:
the point of arrival in the cortex for information from the retina; also called the *striate cortex* and *Brodmann area 17*.

priming:
an experimental technique, particularly useful in studying consciousness, where a stimulus (often not consciously perceived) influences performance on subsequent tasks.

principle of cohesion:
principle of infant folk physics according to which two surfaces are part of the same object if and only if they are in contact.

principle of contact:
principle of infant folk physics according to which only surfaces that are in contact can move together.

principle of continuity:
principle of infant folk physics according to which objects can only move on a single continuous path through space – time.

principle of solidity:
principle of infant folk physics according to which there cannot be more than one object in a place at one time.

prior probability:
in applying Bayes ' s rule, this is the probability assigned to the hypothesis before taking the evidence into account.

prisoner ' s dilemma:
any social exchange or interaction between two players where a player benefits most if she defects while her opponent cooperates and suffers most when she cooperates and her opponent defects. If each player is rational and works backward from what her opponent might do, she will always reason that the best choice is to defect.

propositional attitude:
a psychological state that can be analyzed into a proposition (e.g., the proposition that it is snowing in St. Louis) and an attitude toward that proposition (e.g., the attitude of belief or the attitude of hope).

propositional calculus:
formal system for exploring the logical relations between formulas built up from symbols for complete propositions using logical operators (such as NOT, OR, and AND).

psychophysics:
the branch of psychology that studies the relationship between physical stimuli and how subjects perceive and discriminate them.

recurrent network:
an artificial neural network that has a feedback loop serving as a memory

of what the hidden layer was doing at the previous time step.

recursive definition:
process for defining a set of objects by starting with a set of base cases and specifying which transformations of objects preserve membership in the set. So, for example, a recursive definition of a well-formed formula in the propositional calculus starts with propositional symbols (the base cases) and indicates which logical operations (e.g., negation) create new formulas from existing formulas.

reduction:
the process of showing how higher-level parts of science (e.g., thermodynamics) can be understood in terms of more basic parts of science (e.g., statistical mechanics).

reinforcement learning:
learning where the feedback is a reward signal (as opposed to the error signal characteristic of supervised learning).

representation:
structure carrying information about the environment. Representations can be physical symbol structures or distributed states of neural networks.

representation learning:
the subfield of machine learning dedicated to designing algorithms that will do their own feature engineering on raw data.

retrograde amnesia:
the loss of memory of events before a brain injury.

robot reply (to the Russian room argument):
a response to John Searle ′s thought experiment that claims that the Russian room is not intelligent because it is incapable of interacting with other Russian speakers rather than because of any gap between syntax and semantics.

Russian room argument:
John Searle ′s thought experiment that attempts to refute the physical symbol system hypothesis by showing that there can be syntactic symbol manipulation without any form of intelligence or understanding.

saccadic eye movements:
quick and unconscious eye movements scanning the visual field.

segregation, principle of:
fundamental principle of neuroscience stating that the cerebral cortex is divided into separate areas with distinct neuronal populations.

selection processor:
mechanism hypothesized by Leslie enabling people to inhibit the default setting of a true belief. It is not until the selection processor is fully in place that children can pass the false-belief task, according to Leslie.

selective attention:

the ability of individuals to orient themselves toward, or process information from, only one stimulus within the environment, to the exclusion of others.

semantic priming:
a priming task in which the priming effect is due to the meaning of words and not due to their phonology (how they are pronounced) or their orthography (how they are spelled).

semantic property:
a property of a representation that holds by virtue of its content, that is, how it represents the world (e.g., a semantic property of the word *apple* is the fact that it represents a crisp and juicy fruit).

shared weights:
feature of artificial neural networks where multiple units in a single layer have the same weights. Makes for more efficient processing.

simulation theory (radical):
the theory that mindreading takes place when people think about the world from another person ' s perspective rather than thinking about the other person ' s psychological states.

simulation theory (standard):
the theory that people are able to reason about the mental states of others and their consequent potential behaviors by inputting " pretend " beliefs and desires into their own decision-making systems.

situated cognition:
situated-cognition theorists complain that traditional cognitive science has focused on disembodied systems that operate in highly simplified and prepackaged environments. They call instead for an approach to cognitive science that takes seriously the fact that cognitive agents are both embodied and situated within a complex environment.

S.M.:
a well-studied female patient who has difficulty processing fear as a result of bilateral damage of the amygdala yet has normal perceptual and cognitive abilities.

sparse connectivity:
feature of artificial neural networks where units in a given layer are only connected to a proper subset of units in the next layer.

spatial resolution:
the degree of spatial detail provided by a particular technique for studying the brain.

state space:
the state space of a system is a geometrical representation of all the possible states that the system can be in. It has as many dimensions as the system has independently varying quantities.

striate cortex:
see *primary visual cortex* .

subcortex:
the part of the brain, popularly called " white matter, " that developed earlier in evolution than the cerebral cortex.

---

subsumption architecture:
architectures in robotics that are built up incrementally from semiautonomous layers. Subsumption architectures (originally proposed by Rodney Brooks) typically exploit direct links between perception and action.

---

supervised learning:
learning (e.g., in neural networks) that involves an explicit error signal.

---

surface structure:
in Chomskyan linguistics, the surface structure of a sentence is given by the actual arrangement of written or spoken lexical items — as opposed to its deep structure.

---

symbol-grounding problem:
the problem of determining how syntactically manipulated symbols gain semantic meaning.

---

synapse:
the site where the end of an axon branch comes close to a dendrite or the cell body of another neuron. This is where signals are transmitted from one neuron to another.

---

systems neuroscience:

the investigation of the function of neural systems, such as the visual system or auditory system.

systems reply (to the Russian room argument):
a response to John Searle ' s thought experiment claiming that the Russian room as a whole understands Russian, even though the person inside the room does not.

temporal resolution:
the degree of temporal detail provided by a particular technique for studying the brain.

theory-of-mind mechanism (TOMM):
a hypothesized cognitive system specialized for attributing propositional attitudes and using those attributions to predict and explain behavior.

threshold:
the minimum amount of activity necessary to initiate the firing of a unit in an artificial neural network.

TIT FOR TAT:
a successful strategy used in social exchanges, such as the prisoner ' s dilemma, whereby a player cooperates with his opponent during the first round and in subsequent rounds copies the action taken by the opponent on the preceding round.

transformational grammar:

a theoretical account of the rules governing how surface structures in natural languages are generated from deep structures.

truth condition:
the state of affairs that makes a particular statement true.

truth rule:
a rule that states the truth condition for a given statement.

Turing machine:
a theoretical model of an abstract computation device that can (according to the Church – Turing thesis) compute any effectively calculable function.

unilateral spatial neglect:
a neurological disorder, typically resulting from damage to the posterior parietal cortex in one hemisphere, in which patients describe themselves as unaware of stimuli in the contralateral half of their visual field.

unsupervised learning:
learning (e.g., in neural networks) where there is no explicit error or reward signal.

utility:
a widely used concept for measuring the value of an action or outcome to an individual. In Bayesianism, rational agents maximize expected utility.

valence:

a dimension from pleasure to displeasure that characterizes an emotion by measuring the attractiveness or aversiveness of a situation.

ventral pathway:
the neural pathway believed to be specialized for visual information relevant to recognizing and identifying objects. This pathway runs from the primary visual cortex to the temporal lobe.

Wason selection task:
experiment developed to test people ' s understanding of conditional reasoning. Subjects are asked to identify the additional information they would need in order to tell whether a given conditional statement is true or false.

well-formed formula:
a string of symbols in a formal language that is legitimately constructed through the formation rules of that language.

Bibliography

Abiri , R. , Borhani , S. , Sellers , E. W. , Jiang , Y. , and Zhao , X. (2019). A comprehensive review of EEG-based brain − computer interface paradigms . *Journal of Neural Engineering* , ***16*** , 011001 . doi: 10.1088/1741-2552/aaf12e

Abreu , R. , Leal , A. , and Figueiredo , P. (2018). EEG-informed fMRI: A review of data analysis methods . *Frontiers in Human Neuroscience* , ***12*** , 29 .

Abu-Akel , A. , and Shamay-Tsoory , S. (2011). Neuroanatomical and neurochemical bases of theory of mind . *Neuropsychologia* , ***49*** , 2971 − 84 .

Adams , F. , and Aizawa , A. (2010). *The Bounds of Cognition* . Oxford : Wiley-Blackwell .

Adolphs , R. (2009). The social brain: Neural basis of social knowledge . *Annual Review of Psychology* , ***60*** , 693 − 716 .

Adolphs , R. (2013). The biology of fear . *Current Biology* , ***23*** , R79 − R93 . doi: 10.1016/j.cub.2012.11.055

Adolphs , R. , and Anderson , D. J. (2018). *The Neuroscience of Emotion: A New Synthesis* . Princeton, NJ : Princeton University Press .

Adolphs , R. , and Tranel , D. (2000). Emotion recognition and the human amygdala . In J. P. Aggleton (ed.), *The Amygdala: A Functional Analysis* . Oxford : Oxford University Press .

Adolphs , R. , Tranel , D. , Damasio , H. , and Damasio , A. (1994). Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala . *Nature* , ***372*** , 669 – 72 .

Aglioti , S. , DeSouza , J. F. X. , and Goodale , M. A. (1995). Size – contrast illusions deceive the eye but not the hand . *Current Biology* , ***5*** , 679 – 85 .

Albright , T. D. (2017). Why eyewitnesses fail . *Proceedings of the National Academy of Sciences* , ***114*** , 7758 – 64 .

Allott , N. , and Smith , N. (2021). Chomsky and Fodor on modularity . In *A Companion to Chomsky* . Hoboken, NJ : John Wiley & Sons, Ltd . doi: 10.1002/9781119598732.ch34

Amaral , D. G. , and Adolphs , R. (2016). *Living without an Amygdala* . New York : Guildford Press .

Anderson , J. A. (2003). McCulloch-Pitts neurons . In L. Nadel (ed.), *Encyclopedia of Cognitive Science* . New York : Nature Publishing Group .

Anderson , J. R. , Bothell , D. , Byrne , M. D. , Douglass , S. , Lebiere , C. , and Qin , Y. (2004). An integrated theory of the mind . *Psychological*

Review , ***4*** , 1036 – 160 .

Anderson , M. L. , Richardson , M. J. , and Chemero , A. (2012). Eroding the boundaries of cognition: Implications of embodiment . *Topics in Cognitive Science* , ***4*** , 717 – 30 .

Apperly , I. A. , Samson , D. , and Humphreys , G. W. (2005). Domain-specificity and theory of mind: Evaluating neuropsychological evidence . *Trends in Cognitive Sciences* , ***9*** , 572 – 7 .

Apperly , I. A. , Samson , D. , Chiavarino , C. , and Humphreys , G. W. (2004). Frontal and temporoparietal lobe contributions to theory of mind: Neuropsychological evidence from a false-belief task with reduced language and executive demands . *Journal of Cognitive Neuroscience* , ***16*** , 1773 – 84 .

Arbib , M. A. (1987). *Brains, Machines, and Mathematics* . New York : Springer .

Arbib , M. A. (2003). *The Handbook of Brain Theory and Neural Networks* . Cambridge, MA : MIT Press .

Arkin , R. C. (1998). *Behavior-Based Robotics* . Cambridge, MA : MIT Press .

Ashby , F. G. (2011). *Statistical Analysis of fMRI Data* . Cambridge, MA : MIT Press .

Aslin , R. N. , Saffran , J. R. , and Newport , E. L. (1998). Computation of conditional probability statistics by 8-month-old infants . *Psychological Science* , ***9*** , 321 – 4 .

Avneon , M. , and Lamy , D. (2019). Do semantic priming and retrieval of stimulus-response associations depend on conscious perception? *Consciousness and Cognition* , ***69*** , 36 – 51 . doi: 10.1016/j.concog.2019.01.010

Awh , E. , Belopolsky , A. V. , and Theeuwes , J. (2012). Top-down versus bottom-up attentional control: A failed theoretical dichotomy . *Trends in Cognitive Sciences* , ***16*** , 437 – 43 .

Baars , B. J. (1988). *A Cognitive Theory of Consciousness* . Cambridge : Cambridge University Press .

Baars , B. J. (2002). The conscious access hypothesis: Origins and recent evidence . *Trends in Cognitive Sciences* , ***6*** , 47 – 52 .

Baars , B. J. , and Gage , N. M. (eds.). (2010). *Cognition, Brain, and Consciousness: An Introduction to Cognitive Neuroscience* (2nd ed.). Burlington, MA : Elsevier .

Baars , B. J. , and Gage , N. M. (2012). *Fundamentals of Cognitive Neuroscience: A Beginner ′ s Guide* . Waltham, MA : Academic Press .

Baillargeon , R. (1986). Representing the existence and the location of hidden objects: Object permanence in 6- and 8-month-old infants . *Cognition* , ***23*** , 21 – 41 .

Baillargeon , R. (1987). Object permanence in 3- and 4-month-old infants . *Developmental Psychology* , ***23*** , 655 – 64 .

Baillargeon , R. , and Carey , S. (2012). Core cognition and beyond: The acquisition of physical and numerical knowledge . In S. Pauen (ed.), *Early Childhood Development and Later Outcome* . Cambridge : Cambridge University Press .

Baillargeon , R. , Li , J. , Gertner , Y. , and Wu , D. (2010). How do infants reason about physical events? In U. Goswami (ed.), *The Wiley-Blackwell Handbook of Childhood Cognitive Development* (2nd ed.). Oxford : Blackwell .

Baillargeon , R. , Scott , R. M. , and He , Z. (2010). False-belief understanding in infants . *Trends in Cognitive Sciences* , ***14*** , 110 – 8 .

Baillet , S. (2017). Magnetoencephalography for brain electrophysiology and imaging . *Nature Neuroscience* , **20** , 327 – 39 .

Ballard , D. H. , and Zhang , R. (2021). The hierarchical evolution in human vision modeling . *Topics in Cognitive Science* , ***13*** , 309 – 28 . doi: 10.1111/tops.12527

Bandettini , P. A. , and Ungerleider , L. G. (2001). From neuron to BOLD: New connections . *Nature Neuroscience* , ***4*** , 864 – 6 .

Baron-Cohen , S. (1995). *Mindblindness: An Essay on Autism and Theory of Mind* . Cambridge, MA : MIT Press .

Baron-Cohen , S. (2009). The empathizing-systemizing (E-S) theory . *Annals of the New York Academy of Sciences* , ***1156*** , 68 – 80 .

Baron-Cohen , S. , Leslie , A. M. , and Frith , U. (1985). Does the autistic child have a " theory of mind " ? *Cognition* , ***21*** , 37 – 46 .

Baron-Cohen , S. , Tager-Flusberg , H. , and Cohen , D. J. (eds.). (2000). *Understanding Other Minds: Perspectives from Developmental Cognitive Neuroscience* . New York : Oxford University Press .

Barone , P. , Corradi , G. , and Gomila , A. (2019). Infants ′ performance in spontaneous-response false belief tasks: A review and meta-analysis . *Infant Behavior and Development* , ***57*** , 101350 . doi: 10.1016/j.infbeh.2019.101350

Barrett , H. C. , and Kurzban , R. (2006). Modularity in cognition: Framing the debate . *Psychological Review* , ***113*** , 628 – 47 .

Barrett , L. F. (2006). Are emotions natural kinds? *Perspectives on Psychological Science* , ***1*** , 28 – 58 .

Bassett , D. S. , and Bullmore , E. (2006). Small-world brain networks . *Neuroscientist* , ***12*** , 512 – 23 .

Bayne , T. (2012). *The Unity of Consciousness* . New York : Oxford University Press .

Beate , S. (2011). Theory of mind in infancy . *Child Development Perspectives* , ***5*** , 39 – 43 .

Bechara , A. , Damasio , A. R. , Damasio , H. , and Anderson , S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex . *Cognition* , ***50*** , 7 – 15 .

Bechara , A. , Damasio , H. , Damasio , A. R. , and Lee , G. P. (1999). Different contributions of the human amygdala and ventromedial prefrontal cortex to decision-making . *Journal of Neuroscience* , ***19*** , 5473 – 81 .

Bechtel , W. (1998). Representations and cognitive explanations: Assessing the dynamicist ′ s challenge in cognitive science . *Cognitive Science* , ***22*** , 295 – 318 .

Bechtel , W. , and Abrahamsen , A. A. (2002). *Connectionism and the Mind: Parallel Processing, Dynamics and Evolution in Networks* . Cambridge, MA : Blackwell .

Berger , T. W. , Song , D. , Chan , R. H. M. , Marmarelis , V. Z. , LaCoss , J. , Wills , J. , … Granacki , J. J. (2012). A hippocampal cognitive prosthesis: Multi-input, multi-output nonlinear modeling and VLSI implementation . *IEEE Transactions on Neural Systems and Rehabilitation Engineering* , **20** , 198 – 211 .

Bermúdez , J. L. (2005). *Philosophy of Psychology: A Contemporary Introduction* . New York : Routledge .

Bermúdez , J. L. (ed.). (2006). *Philosophy of Psychology: Contemporary Readings* . London : Routledge .

Bermúdez , J. L. (2009). *Decision Theory and Rationality* . Oxford : Oxford University Press .

Bernstein , I. H. , Bissonnette , V. , Vyas , A. , and Barclay , P. (1989). Semantic priming: Subliminal perception or context? *Perception and Psychophysics* , ***45*** , 153 – 61 .

Bernth , J. E. , Ho , V. A. , and Liu , H. (2018). Morphological computation in haptic sensation and interaction: from nature to robotics . *Advanced Robotics* , ***32*** , 340 – 62 .

Berti , A. , and Rizzolatti , G. (1992). Visual processing without awareness: Evidence from unilateral neglect . *Journal of Cognitive Neuroscience* , ***4*** , 345 – 51 .

Berwick , R. C. , Pietroski , P. , Yankama , B. , and Chomsky , N. (2011). Poverty of the stimulus revisited . *Cognitive Science* , ***35*** , 1207 – 42 .

Bickle , J. (2006). Reducing mind to molecular pathways: Explicating the reductionism implicit in current cellular and molecular neuroscience . *Synthese* , ***151*** , 411 – 34 .

Bisiach , E. , and Luzzatti , C. (1978). Unilateral neglect of representational space . *Cortex* , ***14*** , 129 – 33 .

Blamire , A. M. , Ogawa , S. , Ugurbil , K. , Rothman , D. , McCarthy , G. , Ellermann , J. M. , … Shulman , R. G. (1992). Dynamic mapping of the human visual cortex by high-speed magnetic resonance imaging . *Proceedings of the National Academy of Sciences* , ***89*** , 11069 – 73 .

Block , N. (ed.). (1981). *Imagery* . Cambridge, MA : MIT Press .

Block , N. (1995 a). The mind as the software of the brain . In D. Osherson , L. Gleitman , S. M. Kosslyn , E. Smith , and R. J. Sternberg (eds.), *An Invitation to Cognitive Science* . Cambridge, MA : MIT Press .

Block , N. (1995 b). On a confusion about the function of consciousness . *Behavioral and Brain Sciences* , ***18*** , 227 – 47 .

Block , N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience . *Behavioral and Brain Sciences* , ***30*** , 481 – 548 .

Block , N. (2011). Perceptual consciousness overflows cognitive access . *Trends in Cognitive Sciences* , ***15*** , 567 – 75 .

Block , N. , Flanagan , O. , and Güzeldere , G. (eds.). (1997). *The Nature of Consciousness: Philosophical Debates* . Cambridge, MA : MIT Press .

Bloom , P. , and German , T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind . *Cognition* , ***77*** , B25 – B31 .

Boden , M. A. (1977). *Artificial Intelligence and Natural Man* . Brighton : Harvester Press .

Boden , M. A. (1990 a). *Escaping from the Chinese room. In The Philosophy of Artificial Intelligence* . Oxford : Oxford University Press .

Boden , M. A. (ed.). (1990 b). *The Philosophy of Artificial Intelligence* . Oxford : Oxford University Press .

Boden , M. A. (2006). *Mind as Machine: A History of Cognitive Science* . Oxford : Oxford University Press .

Bornhofen , C. , and McDonald , S. (2008). Emotion perception deficits following traumatic brain injury: A review of the evidence and rationale for

intervention . *Journal of the International Neuropsychological Society* , ***14*** , 511 – 25 .

Boucher , J. (1996). What could possibly cause autism? In P. Carruthers and P. K. Smith (eds.), *Theories of Theory of Mind* . Cambridge : Cambridge University Press .

Bowers , J. S. (2009). On the biological plausibility of grandmother cells: Implications for neural network theories in psychology and neuroscience . *Psychological Review* , ***16*** , 220 – 51 .

Bremner , G. J. (1994). *Infancy* . Oxford : Wiley-Blackwell .

Bressler , S. L. , Tang , W. , Sylvester , C. M. , Shulman , G. L. , and Corbetta , M. (2008). Top-down control of human visual cortex by frontal and parietal cortex in anticipatory visual spatial attention . *Journal of Neuroscience* , ***28*** , 10056 – 61 .

Brewer , J. B. , Zhao , Z. , Desmond , J. E. , Glover , G. H. , and Gabrieli , J. D. E. (1998). Making memories: Brain activity that predicts how well visual experience will be remembered . *Science* , ***281*** , 1185 – 7 .

Broadbent , D. E. (1954). The role of auditory localization in attention and memory span . *Journal of Experimental Psychology* , ***47*** , 191 – 6 .

Broadbent , D. E. (1958). *Perception and Communication* . London : Pergamon Press .

Bromberg-Martin , E. S. , and Monosov , I. E. (2020). Neural circuitry of information seeking . *Current Opinion in Behavioral Sciences* , ***35*** , 62 - 70 . https://doi.org/10.1016/j.cobeha.2020.07.006

Brook , A. (2007). *The Prehistory of Cognitive Science* . New York : Palgrave Macmillan .

Brooks , R. (1991). Intelligence without representation . *Artificial Intelligence* , ***47*** , 139 – 59 . Reprinted in Haugeland, J. (ed.). (1997). *Mind Design II: Philosophy, Psychology, Artificial Intelligence* . Cambridge, MA: MIT Press.

Brooks , R. (1999). *Cambrian Intelligence: The Early History of the New AI* . Cambridge, MA : MIT Press .

Broome , J. (1991). Utility . *Economics and Philosophy* , ***7*** , 1 – 12 .

Buckley , R. F. , Schultz , A. P. , Hedden , T. , Papp , K. V. , Hanseeuw , B. J. , Marshall , G. , … Chhatwal , J. P. (2017). Functional network integrity presages cognitive decline in preclinical Alzheimer disease . *Neurology* , ***89*** , 29 – 37 .

Buckner , R. L. (1998). Event-related fMRI and the hemodynamic response . *Human Brain Mapping* , ***6*** , 373 – 7 .

Bullmore , E. , and Sporns , O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems . *Nature Reviews Neuroscience* , ***10*** , 186 – 98 .

Burkart , J. M. , Schubiger , M. N. , and van Schaik , C. P. (2017). The evolution of general intelligence . *Behavioral and Brain Sciences* , ***40*** , e195 . doi: 10.1017/S0140525X16000959

Busey , T. A. , and Loftus , G. R. (2007). Cognitive science and the law . *Trends in Cognitive Sciences* , ***11*** , 111 – 7 .

Cacchione , T. (2013). The foundations of object permanence: does perceived cohesion determine infants ′ appreciation of the continuous existence of material objects? *Cognition* , ***128*** , 397 – 406 .

Cadieu , C. F. , Hong , H. , Yamins , D. L. K. , Pinto , N. , Ardila , D. , Solomon , E. A. , … DiCarlo , J. J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition . *PLoS Computational Biology* , ***10*** , e1003963 .

Calvo , P. , and Symons , J. (2014). *The Architecture of Cognition: Rethinking Fodor and Pylyshyn ′ s Systematicity Challenge* . Cambridge, MA : MIT Press .

Carey S. (2009). *The Origin of Concepts* . Oxford : Oxford University Press .

Carey , S. , and Spelke , E. S. (1996). Science and core knowledge . *Philosophy of Science* , ***63*** , 515 – 33 .

Carrasco , M. (2011). Visual attention: The past 25 years . *Vision Research* , ***51*** , 1484 – 535 .

Carrasco , M. (2018). How visual spatial attention alters perception . *Cognitive Processing* , **19** (Suppl. 1), 77 – 88 .

Carrington , S. J. , and Bailey , A. J. (2009). Are there theory of mind regions in the brain? A review of the neuroimaging literature . *Human Brain*

Mapping , ***30*** , 2313 – 35 .

Carruthers , P. (2000). *Phenomenal Consciousness* . Cambridge : Cambridge University Press .

Carruthers , P. (2006). *The Architecture of the Mind* . Cambridge : Cambridge University Press .

Carruthers , P. (2008 a). On Fodor-fixation, flexibility, and human uniqueness: A reply to Cowie, Machery, and Wilson . *Mind and Language* , ***23*** , 293 – 303 .

Carruthers , P. (2008 b). Precis of *The Architecture of the Mind: Massive Modularity and the Flexibility of Thought* . *Mind and Language* , ***23*** , 257 – 62 .

Carruthers , P. (2013). Mindreading in infancy . *Mind and Language* , ***28*** , 141 – 72 .

Carruthers , P. , and Smith , P. K. (eds.). (1996). *Theories of Theory of Mind* . Cambridge : Cambridge University Press .

Chalmers , D. (1995). Facing up to the problem of consciousness . *Journal of Consciousness Studies* , ***2*** , 200 – 19 .

Chalmers , D. (1996). *The Conscious Mind* . Oxford : Oxford University Press .

Chalmers , D. J. (2013). How can we construct a science of consciousness? *Annals of the New York Academy of Sciences* , ***1303*** , 25 – 35 .

Charpac , S. , and Stefanovic , B. (2012). Shedding light on the BOLD fMRI response . *Nature Methods* , ***9*** , 547 – 9 .

Chater , N. , and Manning , C. D. (2006). Probabilistic models of language processing and acquisition . *Trends in Cognitive Sciences* , ***10*** , 335 – 44 .

Chater , N. , and Oaksford , M. (2008). *The Probabilistic Mind: Prospects for Bayesian Cognitive Science* . Oxford : Oxford University Press .

Chater , N. , Oasksford , M. , Hahn , U. , and Heit , E. (2010). Bayesian models of cognition . *Wiley Interdisciplinary Reviews: Cognitive Science* , ***1*** , 811 – 23 .

Chelazzi , L. , and Corbetta , M. (2000). Cortical mechanisms of visuospatial attention in the primate brain . In M. S. Gazzaniga (ed.), *The New Cognitive Neurosciences* (2nd ed.). Cambridge, MA : MIT Press .

Chemero , A. (2009). *Radical Embodied Cognitive Science* . Cambridge, MA : MIT Press .

Cherry , E. C. (1953). Some experiments on the recognition of speech, with one and two ears . *Journal of the Acoustical Society of America* , ***25*** , 975 – 9 .

Chhatwal , J. P. , Schultz , A. P. , Johnson , K. , Benzinger , T. L. S. , Jack , C. , Ances , B. M. , … Sperling , R. A. (2013). Impaired default network functional connectivity in autosomal dominant Alzheimer disease . *Neurology* , ***81*** , 736 – 44 .

Chomsky , N. (1957). *Syntactic Structures* . Gravenhage : Mouton .

Chomsky , N. (1959). A review of B. F. Skinner ′ s *Verbal Behavior* . *Language* , ***35*** , 26 – 58 .

Chomsky , N. (1968). *Language and Mind* . New York : Harper and Row .

Chomsky , N. (1980 a). *Rules and Representations* . New York : Columbia University Press .

Chomsky , N. A. (1980 b). Rules and representations . *Behavioral and Brain Sciences* , ***3*** , 1 – 61 .

Chow , R. , Rabi , R. , Paracha , S. , Hasher , L. , Anderson , N. D. , and Alain , C. (2022). Default mode network and neural phase synchronization in healthy aging: A resting state EEG study . *Neuroscience* , ***485*** , 116 – 28 . doi: 10.1016/j.neuroscience.2022.01.008

Christiansen , M. H. , and Chater , N. (2001). *Connectionist Psycholinguistics* . Westport, CT : Ablex .

Chun , M. M. , Golomb , J. D. , and Turk-Browne , N. B. (2011). A taxonomy of external and internal attention . *Annual Review of Psychology* , ***62*** , 73 – 101 .

Churchland , P. M. (1990 a). *On the nature of theories: A neurocomputational perspective. In A Neurocomputational Perspective: The Nature of Mind and the Structure of Science* . Cambridge, MA : MIT Press .

Churchland , P. M. (1990 b). Cognitive activity in artificial neural networks . In N. Block and D. Osherson (eds.), *Invitation to Cognitive Science* . Cambridge, MA : MIT Press . Reprinted in Cummins, R., and Cummins, D.

D. (2000). *Minds, Brains, and Computers: The Foundations of Cognitive Science: An Anthology* . Malden, MA: Blackwell.

Churchland , P. M. (2007). *Neurophilosophy at Work* . Cambridge : Cambridge University Press .

Churchland , P. M. , and Churchland , P. S. (1990). Could a machine think? *Scientific American* , ***262*** , 32 – 7 .

Churchland , P. S. (1986). *Neurophilosophy: Toward a Unified Science of the Mind/Brain* . Cambridge, MA : MIT Press .

Churchland , P. S. , and Sejnowski , T. J. (1992). *The Computational Brain* . Cambridge, MA : MIT Press .

Cichy , R. M. , Khosla , A. , Pantazis , D. , Torralba , A. , and Oliva , A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence . *Scientific Reports* , ***6*** , 27755. doi: 10.1038/srep27755

Clancey , W. J. (1997). *Situated Cognition: On Human Knowledge and Computer Representations* . Cambridge : Cambridge University Press .

Clark , A. (1989). *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing* . Cambridge, MA : MIT Press .

Clark , A. (1993). *Associative Engines: Connectionism, Concepts, and Representational Change* . Cambridge, MA : MIT Press .

Clark , A. (1997). *Being There: Putting Brain, Body, and World Together Again* . Cambridge, MA : MIT Press .

Clark , A. (1998). Time and mind . *Journal of Philosophy* , ***95*** , 354 – 76 .

Clark , A. (2001). *Mindware: An Introduction to the Philosophy of Cognitive Science* . New York : Oxford University Press .

Clark , A. (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension* . New York : Oxford University Press .

Clark , A. (2011). Précis of *Supersizing the Mind: Embodiment, Action, and Cognitive Extension* . *Philosophical Studies* , ***152*** , 413 – 6 .

Clark , A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind* . Oxford : Oxford University Press .

Clarke , S. (2021). Cognitive penetration and informational encapsulation: Have we been failing the module? *Philosophical Studies* , ***178*** , 2599 – 620 .

Clearfield , M. W. , Dineva , E. , Smith , L. B. , Diedrich , F. J. , and Thelen , E. (2009). Cue salience and infant perseverative reaching: Tests of the dynamic field theory . *Developmental Science* , ***12*** , 26 – 40 .

Colby , C. L. , and Goldberg , M. E. (1999). Space and attention in parietal cortex . *Annual Review of Neuroscience* , ***22*** , 319 – 49 .

Coleman , S. (2019). *The Knowledge Argument* . Cambridge University Press .

Colombo , M. , Elkin , L. , and Hartmann , S. (2021). Being realist about Bayes, and the predictive processing theory of mind . *British Journal for*

the Philosophy of Science , ***72*** , 185 – 220 .

Constant , A. , Clark , A. , and Friston , K. J. (2021). Representation wars: Enacting an armistice through active inference . *Frontiers in Psychology* , ***11*** . doi: 10.3389/fpsyg.2020.598733

Cook , R. , Bird , G. , Catmur , C. , Press , C. , and Heyes , C. (2014). Mirror neurons: From origin to function . *Behavioral and Brain Sciences* , ***37*** , 177 – 92 .

Cook , V. J. , and Newson , M. (2007). *Chomsky ' s Universal Grammar: An Introduction* (3rd ed.). Oxford : Blackwell .

Cooper , L. A. , and Shepard , R. N. (1973). The time required to prepare for a rotated stimulus . *Memory and Cognition* , ***1*** , 246 – 50 .

Copeland , J. G. (1993). *Artificial Intelligence: A Philosophical Introduction* . Oxford : Blackwell .

Corbetta , M. , and Shulman , G. L. (2011). Spatial neglect and attention networks . *Annual Review of Neuroscience* , ***34*** , 569 – 99 .

Cosmides , L. , Barrett , H. C. , and Tooby , J. (2010). Adaptive specializations, social exchange, and the evolution of human intelligence . *Proceedings of the National Academy of Sciences USA* , ***107*** , 9007 – 14 .

Cosmides , L. , and Tooby , J. (1994). Origins of domain-specificity: The evolution of functional organization . In L. A. Hirschfeld and S. F. Gelman (eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture* . Cambridge : Cambridge University Press . Reprinted in Bermúdez, J. L.

(ed.). (2006). *Philosophy of Psychology: Contemporary Readings* . London: Routledge.

Cosmides , L. , and Tooby , J. (2013). Evolutionary psychology: New perspectives on cognition and motivation . *Annual Review of Psychology* , **64** , 201 – 29 .

Cowie , F. (1999). *What ′ s Within? Nativism Reconsidered* . Oxford : Oxford University Press .

Cowie , F. (2008). Us, them and it: Modules, genes, environments and evolution . *Mind and Language* , ***23*** , 284 – 92 .

Crane , T. (2003). *The Mechanical Mind: A Philosophical Introduction to Minds, Machines, and Mental Representation* . London : Routledge .

Craver , C. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience* . New York : Oxford University Press .

Crick , F. , and Koch , C. (2003). A framework for consciousness . *Nature Neuroscience* , ***6*** , 119 – 26 .

Croker , V. , and McDonald , S. (2005). Recognition of emotion from facial expression following traumatic brain injury . *Brain Injury* , ***19*** , 787 – 99 .

Cummins , R. , and Cummins , D. D. (2000). *Minds, Brains, and Computers: The Foundations of Cognitive Science: An Anthology* . Malden, MA : Blackwell .

Cutland , N. J. (1980). *Computability: An Introduction to Recursive Function Theory* . Cambridge : Cambridge University Press .

Dale , A. M. , and Buckner , R. L. (1997). Selective averaging of rapidly presented individual trials using fMRI . *Human Brain Mapping* , ***5*** , 329 – 40 .

Damasio , A. (1994). *Descartes ' s Error: Emotion, Reason, and the Human Brain* . New York : Putnam/Grosset .

Damasio , H. , Grabowski , T. , Frank , R. , Galaburda , A. M. , and Damasio , A. R. (1994). The return of Phineas Gage: Clues about the brain from the skull of a famous patient . *Science* , ***264*** , 1102 – 5 . doi: 10.1126/science.8178168

Darby , R. R. , Horn , A. , Cushman , F. , and Fox , M. D. (2018). Lesion network localization of criminal behavior . *Proceedings of the National Academy of Sciences* , ***115*** , 601 – 6 . doi: 10.1073/pnas.1706587115

Darwin , C. (1890). *The Expression of the Emotions in Man and Animals* (F. Darwin , ed.) (2nd ed.). London : John Murray .

Davies , M. , and Stone , T. (eds.). (1995 a). *Folk Psychology* . Oxford : Blackwell .

Davies , M. , and Stone , T. (eds.). (1995 b). *Mental Simulation* . Oxford : Blackwell .

Davis , M. (1992). The role of the amygdala in fear and anxiety . *Annual Review of Neuroscience* , ***15*** , 353 – 75 .

Davis , M. (2000). *The Universal Computer: The Road from Leibniz to Turing* . New York : Norton .

Davis , M. (2001). *Engines of Logic: Mathematicians and the Origin of the Computer* . New York : Norton .

Davis , S. (1993). *Connectionism: Theory and Practice* . New York : Oxford University Press .

Daw , N. D. , O ′ Doherty , J. P. , Dayan , P. , Seymour , B. , and Dolan , R. J. (2006). Cortical substrates for exploratory decisions in humans . *Nature* , ***441*** , 876 – 9 .

Dawkins , R. (1979). Twelve misunderstandings of kin selection . *Zeitschrift für Tierpsychologie* , ***51*** , 184 – 200 .

Dawson , M. R. W. (1998). *Understanding Cognitive Science* . Oxford : Blackwell .

Dawson , M. R. W. (2004). *Minds and Machines: Connectionism and Psychological Modeling* . Oxford : Blackwell .

Dawson , M. R. W. (2005). *Connectionism: A Hands-On Approach* . Oxford : Blackwell .

Dayan , P. , and Abbott , L. F. (2005). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems* . Cambridge, MA : MIT Press .

de Jong , M. C. , Vansteensel , M. J. , van Ee , R. , Leijten , F. S. S. , Ramsey , N. F. , Dijkerman , H. C. , … Knapen , T. (2020). Intracranial recordings reveal unique shape and timing of responses in human visual cortex during illusory visual events . *Current Biology* , ***30*** , 3089 – 100 .e3084.

Dehaene , S. , and Changeux , J. (2011). Experimental and theoretical approaches to conscious processing . *Neuron* , ***70*** , 200 – 27 .

Dehaene , S. , Changeux , J. , Naccache , L. , Sackur , J. , and Sergent , C. (2006). Conscious, preconscious, and subliminal processing: A testable taxonomy . *Trends in Cognitive Sciences* , ***10*** , 204 – 11 .

Dehaene , S. , Charles , L. , King , J.-R. , and Marti , S. (2014). Toward a computational theory of conscious processing . *Current Opinion in Neurobiology* , ***25*** , 76 – 84 .

Dehaene , S. , Kerszberg , M. , and Changeux , J. P. (1998). A neuronal model of a global workspace in effortful cognitive tasks . *Proceedings of the National Academy of Sciences USA* , ***95*** , 14529 – 34 .

Dehaene , S. , and Naccache , L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework . *Cognition* , ***79*** , 1 – 37 .

Dennett , D. C. (1969). *Content and Consciousness* . London : Routledge & Kegan Paul .

Dennett , D. C. (1991). *Consciousness Explained* . Boston : Little, Brown .

Deshpande , G. , and Hu , X. (2012). Investigating effective brain connectivity from fMRI data: Past findings and current issues with reference to Granger causality analysis . *Brain Connectivity* , ***2*** , 235 – 45 .

Dimov , C. , Khader , P. H. , Marewski , J. N. , and Pachur , T. (2020). How to model the neurocognitive dynamics of decision making: A methodological primer with ACT-R . *Behavior Research Methods* , ***52*** , 857 – 80 .

Doerig , A. , Schurger , A. , and Herzog , M. H. (2021). Hard criteria for empirical theories of consciousness . *Cognitive Neuroscience* , ***12*** , 41 – 62 . doi: 10.1080/17588928.2020.1772214

Dreyfus , H. L. (1977). *Artificial Intelligence and Natural Man* . New York : Basic Books .

Driver , J. (2001). A selective review of selective attention research from the past century . *British Journal of Psychology* , ***92*** , 53 – 78 .

Driver , J. , and Mattingly , J. B. (1998). Parietal neglect and visual awareness . *Nature Neuroscience* , ***1*** , 17 – 22 .

Driver , J. , and Vuilleumier , P. (2001). Perceptual awareness and its loss in unilateral neglect and extinction . *Cognition* , ***79*** , 39 – 88 .

Du , K.-L. , and Swamy , M. N. S. (2019). Perceptrons . In K.-L. Du and M. N. S. Swamy (eds.), *Neural Networks and Statistical Learning* . London : Springer .

Dukes , D. , Abrams , K. , Adolphs , R. , Ahmed , M. E. , Beatty , A. , Berridge , K. C. , … Sander , D. (2021). The rise of affectivism . *Nature Human Behaviour* , ***5*** , 816 – 20 .

Dunn , B. D. , Dalgleish , T. , and Lawrence , A. D. (2006). The somatic marker hypothesis: A critical evaluation . *Neuroscience & Biobehavioral Reviews* , ***30*** , 239 – 271 .

Ekman , P. (1992). An argument for basic emotions . *Cognition and Emotion* , ***63*** , 169 – 200 .

Ekman , P. , and Friesen , W. V. (1971). Constants across cultures in the face and emotion . *Journal of Personality and Social Psychology* , ***17*** , 124 – 9 .

Ekstrom , A. (2010). How and when the fMRI BOLD signal relates to underlying neural activity: The danger in dissociation . *Brain Research Review* , **62** , 233 – 44 .

Eliasmith , C. (1996). The third contender: A critical examination of the dynamicist theory of cognition . *Philosophical Psychology* , ***9*** , 441 – 63 .

Elliott , M. H. (1928). The effect of change or reward on the maze performance of rats . *University of California Publications in Psychology* , ***4*** , 19 – 30 .

Elman , J. L. (2005). Connectionist models of cognitive development: Where next? *Trends in Cognitive Sciences* , ***9*** , 111 – 7 .

Elman , J. L. , Bates , E. A. , Johnson , M. H. , and Karmiloff-Smith , A. (1996). *Rethinking Innateness: A Connectionist Perspective on Development* . Cambridge, MA : MIT Press .

Ernst , M. O. , and Banks , M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion . *Nature* , ***415*** , 429 – 33 .

Fang , F. , and He , S. (2005). Cortical responses to invisible objects in the human dorsal and ventral pathways . *Nature Neuroscience* , ***10*** , 1380 – 5 .

Favela , L. H. (2020). Dynamical systems theory in cognitive science and neuroscience . *Philosophy Compass* , ***15*** , e12695 .

Favela , L. H. (2021). The dynamical renaissance in neuroscience . *Synthese* , ***199*** , 2103 – 27 .

Feinstein , J. S. , Adolphs , R. , Damasio , A. , and Tranel , D. (2011). The human amygdala and the induction and experience of fear . *Current Biology* , ***21*** , 34 – 8 .

Feinstein , J. S. , Buzza , C. , Hurlemann , R. , Follmer , R. L. , Dahdaleh , N. S. , Coryell , W. H. , … Wemmie , J. A. (2013). Fear and panic in humans with bilateral amygdala damage . *Nature Neuroscience* , ***16*** , 270 – 2 .

Felleman , D. J. , and Van Essen , D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex . *Cerebral Cortex* , ***1*** , 1 – 47 .

Fiez , J. A. , Balota , D. A. , Raichle , M. E. , and Petersen , S. E. (1999). Effects of lexicality, frequency, and spelling-to-sound consistency on the functional anatomy of reading . *Neuron* , ***24*** , 205 – 18 . doi: 10.1016/S0896-6273(00)80833-8

Finkbeiner , M. , and Forster , K. I. (2008). Attention, intention and domain-specific processing . *Trends in Cognitive Sciences* , ***12*** , 59 – 64 .

Flanagan , O. J. (1991). *The Science of the Mind* . Cambridge, MA : MIT Press .

Fletcher-Watson , S. , McConnell , F. , Manola , E. , and McConachie , H. (2014). Interventions based on the theory of mind cognitive model for autism spectrum disorder (ASD) . *Cochrane Database Systematic Reviews* , ***3*** .

Fodor , J. (1975). *The Language of Thought* . Cambridge, MA : Harvard University Press .

Fodor , J. (1983). *The Modularity of Mind* . Cambridge, MA : MIT Press .

Fodor , J. (1985). Precis of *The Modularity of Mind* . *Behavioral and Brain Sciences* , ***1*** , 1 – 5 .

Fodor , J. (1987). *Psychosemantics* . Cambridge, MA : MIT Press .

Fodor , J. (2000). *The Mind Doesn ′ t Work That Way: The Scope and Limits of Computational Psychology* . Cambridge, MA : MIT Press .

Fodor , J. (2008). *LOT 2: The Language of Thought Revisited* . Oxford : Oxford University Press .

Fodor , J. , and Pylyshyn , Z. (1988). Connectionism and cognitive architecture: A critical analysis . *Cognition* , ***28*** , 3 – 71 .

Fontaine , J. R. J. , Scherer , K. R. , Roesch , E. B. , and Ellsworth , P. C. (2007). The world of emotions is not two-dimensional . *Psychological Science* , ***18*** , 1050 – 7 . doi: 10.1111/j.1467-9280.2007.02024.x

Fox , A. S. , Lapate , R. C. , Shackman , A. J. , and Davidson , R. J. (2018). *The Nature of Emotion: Fundamental Questions* . Oxford : Oxford University Press .

Fox , E. (2018). Perspectives from affective science on understanding the nature of emotion . *Brain and Neuroscience Advances* , ***2*** . doi: 10.1177/2398212818812628

Frankish , K. , and Ramsey , W. (eds.). (2012). *The Cambridge Handbook of Cognitive Science* . Cambridge : Cambridge University Press .

Franklin , S. (1995). *Artificial Minds* . Cambridge, MA : MIT Press .

Franz , V. H. , Gegenfurtner , K. R. , Bülthoff , H. H. , and Fahle , M. (2000). Grasping visual illusions: No evidence for a dissociation between perception and action . *Psychological Science* , ***11*** , 20 – 5 .

Freyberg , J. , Robertson , C. E. , and Baron-Cohen , S. (2015). Reduced perceptual exclusivity during object and grating rivalry in autism . *Journal of Vision* , ***15*** , 11. doi: 10.1167/15.13.11

Friedenberg , J. , and Silverman , G. (2006). *Cognitive Science: An Introduction to the Study of Mind* . Thousand Oaks, CA : Sage .

Friedman , O. , and Leslie , A. M. (2007). The conceptual underpinnings of pretense: Pretending is not " behaving-as-if. " *Cognition* , **105** , 103 – 24 .

Friedman , O. , Neary , K. R. , Burnstein , C. L. , and Leslie , A. M. (2010). Is young children ′ s recognition of pretense metarepresentational or merely behavioral? Evidence from 2- and 3-year-olds ′ understanding of pretend sounds and speech . *Cognition* , **115** , 314 – 19 .

Friston , K. , Moran , R. , and Seth , A. K. (2013). Analysing connectivity with Granger causality and dynamic causal modelling . *Current Opinion in Neurobiology* , ***23*** , 172 – 8 .

Friston , K. J. (2011). Functional and effective connectivity: A review . *Brain Connectivity* , ***1*** , 13 – 36 .

Frith , C. , and Frith , U. (2012). Mechanisms of social cognition . *Annual Review of Psychology* , ***63*** , 287 – 313 .

Gallistel , C. R. (1990). *The Organization of Learning* . Cambridge, MA : MIT Press .

Garcia-Aunon , P. , del Cerro , J. , and Barrientos , A. (2019). Behavior-based control for an aerial robotic swarm in surveillance missions . *Sensors* , ***19*** , 4584. doi: 10.3390/s19204584

Gardner , H. (1985). *The Mind ' s New Science: A History of the Cognitive Revolution* . New York : Basic Books .

Gazzaniga , M. S. (ed.). (1995). *The New Cognitive Neurosciences* (1st ed.). Cambridge, MA : MIT Press .

Gazzaniga , M. S. (ed.). (2000). *The New Cognitive Neurosciences* (2nd ed.). Cambridge, MA : MIT Press .

Gazzaniga , M. S. (ed.). (2004). *The New Cognitive Neurosciences* (3rd ed.). Cambridge, MA : MIT Press .

Gazzaniga , M. S. (ed.). (2009). *The New Cognitive Neurosciences* (4th ed.). Cambridge, MA : MIT Press .

Gazzaniga , M S. , Halpern , T. , and Heatherton , D. (2011). *Psychological Science* (4th ed.). New York : Norton .

Gazzaniga , M. S. , Ivry , R. B. , and Mangun , G. R. (2013). *Cognitive Neuroscience: The Biology of the Mind* (4th ed.). New York : Norton .

Gazzaniga , M. S. , and Mangun , G. (eds.). (2014). *The New Cognitive Neurosciences* (5th ed.). Cambridge, MA : MIT Press .

Glass , A. L. (2016). *Cognition: A Neuroscience Approach* . Cambridge : Cambridge University Press .

Glasser , M. F. , Smith , S. M. , Marcus , D. S. , Andersson , J. L. R. , Auerbach , E. J. , Behrens , T. E. J. , … Van Essen , D. C. (2016). The

Human Connectome Project ′ s neuroimaging approach . *Nature Neuroscience* , ***19*** , 1175 .

Gleitman , H. , Fridlund , J. , and Reisberg , D. (2010). *Psychology* (8th ed.). New York : Norton .

Glimcher , P. W. (2003). *Decisions, Uncertainty, and the Brain: The Science of Neuroeconomics* . Cambridge MA : MIT Press .

Glimcher , P. W. (2011) Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis . *Proceedings of the National Academy of Sciences USA* , ***108*** , 15647 – 54 .

Glimcher , P. W. , and Fehr , E. (2014). *Neuroeconomics* . Cambridge MA : Academic Press .

Glover , S. R. , and Dixon , P. (2001). Dynamic illusion effects in a reaching task: Evidence for separate visual representations in the planning and control of reaching . *Journal of Experimental Psychology: Human Perception and Performance* , ***27*** , 560 – 72 .

Goense , J. , Whittingstall , K. , and Logothetis , N. K. (2012). Neural and BOLD responses across the brain . *WIREs Cognitive Science* , ***3*** , 75 – 86 .

Goldman , A. (2006). *Simulating Minds* . New York : Oxford University Press .

Goldstein , T. R. , and Lerner , M. D. (2018). Dramatic pretend play games uniquely improve emotional control in young children . *Developmental Science* , ***21*** , e12603. doi: 10.1111/desc.12603

Goodale , M. A. , and Milner , A. D. (2013). *Sight Unseen* (2nd ed.). New York : Oxford University Press .

Goodfellow , I. , Bengio , Y. , and Courville , A. (2016). *Deep Learning* . Cambridge MA : MIT Press .

Gopnik , A. , and Meltzoff , A. N. (1997). *Words, Thoughts, and Theories* . Cambridge, MA : MIT Press .

Gordon , R. (1986). Folk psychology as simulation . *Mind and Language* , ***1*** , 158 – 71 .

Gorman , R. P. , and Sejnowski , T. J. (1988). Analysis of hidden units in a layered network trained to identify sonar targets . *Neural Networks* , ***1*** , 75 – 89 .

Grainger , J. , and Jacobs , A. M. (1998). *Localist Connectionist Approaches to Human Cognition* . Mahwah, NJ : Lawrence Erlbaum .

Graule , M. A. , Chirarattananon , P. , Fuller , S. B. , Jafferis , N. T. , Ma , K. Y. , Spenko , M. , … Wood , R. J. (2016). Perching and takeoff of a robotic insect on overhangs using switchable electrostatic adhesion . *Science* , ***352*** , 978 – 82 .

Greenwald , A. G. , Draine , S. C. , and Abrams , R. L. (1996). Three cognitive markers of unconscious semantic activation . *Science* , ***273*** , 1699 – 702 .

Griffiths , T. L. , Kemp , C. , and Tenenbaum , J. B. (2008). Bayesian models of cognition . In R. Son (ed.), *Cambridge Handbook of*

Computational Cognitive Modeling . Cambridge : Cambridge University Press .

Griggs , R. A. , and Cox , J. R. (1982). The elusive thematic materials effect in the Wason selection task . *British Journal of Psychology* , ***73*** , 407 – 20 .

Gross , C. T. , and Canteras , N. S. (2012). The many paths to fear . *Nature Reviews Neuroscience* , ***13*** , 651 – 8 . doi: 10.1038/nrn3301

Gupta , R. , Koscik , T. R. , Bechara , A. , and Tranel , D. (2011). The amygdala and decision-making . *Neuropsychologia* , ***49*** , 760 – 6 . doi: 10.1016/j.neuropsychologia.2010.09.029

Hacking , I. (2001). *An Introduction to Probability and Inductive Logic* . Cambridge : Cambridge University Press .

Hadley , R. F. (2000). Cognition and the computational power of connectionist networks . *Connection Science* , ***12*** , 95 – 110 .

Hampson , R. E. , Song , D. , Robinson , B. S. , Fetterhoff , D. , Dakos , A. S. , Roeder , B. M. , … Deadwyler , S. A. (2018). Developing a hippocampal neural prosthetic to facilitate human memory encoding and recall . *Journal of Neural Engineering* , ***15*** , 036014.

Hardy-Vallée , B. (2007). Decision-making: A neuroeconomic perspective . *Philosophy Compass* , ***2*** , 939 – 53 .

Harlow , J. M. (1868). Recovery from the passage of an iron bar through the head. *Publications of the Massachusetts Medical Society* , *2* , 327 – 47 . *Reprinted in History of Psychiatry* , ***4*** , 274 – 81 .

Harnad , S. (1990). The symbol-grounding problem . *Physica D* , ***42*** , 335 – 46 .

Hasbach , J. D. , and Bennewitz , M. (2021). The design of self-organizing human – swarm intelligence. *Adaptive Behavior. Advance online publication.* doi: 10.1177/10597123211017550

Hassabis , D. , Kumaran , D. , Summerfield , C. , and Botvinick , M. (2017). Neuroscience-inspired artificial intelligence . *Neuron* , ***95*** , 245 – 58 .

Haugeland , J. (1985). *Artificial Intelligence: The Very Idea* . Cambridge, MA : MIT Press .

Haugeland , J. (1997). *Mind Design II: Philosophy, Psychology, Artificial Intelligence* . Cambridge, MA : MIT Press .

Heal , J. (1986). Replication and functionalism . In J. Butterfield (ed.), *Language, Mind and Logic* . Cambridge : Cambridge University Press .

Hebb , D. O. (1949). *The Organization of Behavior: A Neuropsychological Theory* . New York : Wiley .

Heeger , D. J. , and Ress , D. (2002). What does fMRI tell us about neuronal activity? *Nature Reviews Neuroscience* , ***3*** , 142 – 51 .

Heil , J. (2004). *Philosophy of Mind: A Guide and Anthology* . New York : Oxford University Press .

Henson , R. (2006). Forward inference using functional neuroimaging: Dissociations versus associations . *Trends in Cognitive Sciences* , ***10*** , 64 –

9 .

Hergenhahn , B. R. , and Henley , T. (2013). *An Introduction to the History of Psychology* . Boston : Cengage Learning .

Hespos , S. J. , and van Marle , K. (2012). Physics for infants: Characterizing the origins of knowledge about objects, substances, and number . *WIREs Cognitive Science* , ***3*** , 19 – 27 .

Heyes , C. (2014). False belief in infancy: A fresh look . *Developmental Science* , ***17*** , 647 – 59 .

Heyes , C. , and Catmur , C. (2022). What happened to mirror neurons? *Perspectives on Psychological Science* , ***17*** , 153 – 68 . doi.: 10.1177/1745691621990638

Hinton , G. E. , McClelland , J. L. , and Rumelhart , D. E. (1986). Distributed representations . In D. E. Rumelhart , J. L. McClelland , and the PDP Research Group (eds.), *Parallel Distributed Processing: Explorations in the Microstructures of Cognition: Vol. 1. Foundations* . Cambridge, MA : MIT Press .

Hirose , S. (1993). *Biologically Inspired Robots: Snake-Like Locomotors and Manipulators* . Oxford : Oxford University Press .

Hirschfeld , L. A. , and Gelman , S. F. (eds.). (1994). *Mapping the Mind: Domain Specificity in Cognition and Culture* . Cambridge : Cambridge University Press .

Hodges , A. (2014). *Alan Turing: The Enigma* . Princeton, NJ : Princeton University Press .

Hofhansel , L. , Weidler , C. , Votinov , M. , Clemens , B. , Raine , A. , and Habel , U. (2020). Morphology of the criminal brain: Gray matter reductions are linked to antisocial behavior in offenders . *Brain Structure and Function* , ***225*** , 2017 – 28 .

Hohwy , J. (2009). The neural correlates of consciousness: New experimental approaches needed? *Consciousness and Cognition* , ***18*** , 428 – 38 .

Hohwy , J. (2013). *The Predictive Mind* . Oxford : Oxford University Press .

Hohwy , J. , Roepstorff , A. , and Friston , K. (2008). Predictive coding explains binocular rivalry: An epistemological review . *Cognition* , ***108*** , 687 – 701 .

Hopfinger , J. B. , Luck , S. J. , and Hillyard , S. A. (2004). Selective attention: Electrophysiological and neuromagnetic studies . In M. Gazzaniga (ed.), *The Cognitive Neurosciences* (3rd ed.). Cambridge, MA : MIT Press .

Horwich , P. (1982). *Probability and Evidence* . Cambridge : Cambridge University Press .

Houghton , G. (2005). *Connectionist Models in Cognitive Psychology* . Oxford : Oxford University Press .

Howard , C. J. , Pole , R. , Montgomery , P. , Woodward , A. , Guest , D. , Standen , B. , … Crowe , E. M. (2020). Visual spatial attention and spatial

working memory do not draw on shared capacity-limited core processes . *Quarterly Journal of Experimental Psychology* , ***73*** , 799 – 818 .

Huang , C.-C. , Rolls , E. T. , Feng , J. , and Lin , C.-P. (2021). *An extended Human Connectome Project multimodal parcellation atlas of the human cortex and subcortical areas* . *Brain Structure and Function* . Advance online publication . doi: 10.1007/s00429-021-02421-6

Huettel , S. A. (2012). Event-related fMRI in cognition . *Neuroimage* , ***62*** , 1152 – 6 .

Humphreys , G. W. , Duncan , J. , and Treisman , A. (eds.). (1999). *Attention, Space, and Action: Studies in Cognitive Neuroscience* . Oxford : Oxford University Press .

Husain , M. , and Nachev , P. (2007). Space and the parietal cortex . *Trends in Cognitive Sciences* , ***11*** , 30 – 6 .

Hutchins , E. (1995). *Cognition in the Wild* . Cambridge, MA : MIT Press .

Iacoboni , M. , and Dapretto , M. (2006). The mirror neuron system and the consequences of its dysfunction . *Nature Reviews Neuroscience* , *7* , 942 – 51 .

Ijspeert , A. J. (2014). Biorobotics: Using robots to emulate and investigate agile locomotion . *Science* , ***346*** , 196 – 203 .

Isac , D. , and Reiss , C. (2013). *I-Language: An introduction to linguistics as cognitive science* (2nd ed.). Oxford : Oxford University Press .

Jackson , F. (1982). Epiphenomenal qualia . *Philosophical Quarterly* , ***32*** , 127 – 36 .

Jackson , F. (1986). What Mary didn ′ t know . *Journal of Philosophy* , ***83*** , 291 – 5 .

Jackson , F. (2003). Mind and illusion . In A. O ′ Hear (ed.), *Minds and Persons: Royal Institute of Philosophy Supplement* . Cambridge : Cambridge University Press .

Jackson , P. (1998). *Introduction to Expert Systems* . Harlow, UK : Addison-Wesley .

Jacob , P. , and Jeannerod , M. (2003). *Ways of Seeing: The Scope and Limits of Visual Cognition* . New York : Oxford University Press .

Jacobs , R. A. , and Kruschke , J. K. (2011). Bayesian learning theory applied to human cognition . *Wiley Interdisciplinary Reviews: Cognitive Science* , ***2*** , 8 – 21 .

Jacobson , S. , and Degotardi , S. (2022). A dynamic systems approach to joint attention in an infant-toddler early childhood centre . *Learning, Culture and Social Interaction* , ***32*** , 100574 .

Jansen , P. A. , and Watter , S. (2012). Strong systematicity through sensorimotor conceptual grounding: An unsupervised, developmental approach to connectionist sentence processing . *Connection Science* , ***24*** , 25 – 55 .

Jeffrey , R. (1983). *The Logic of Decision* . Chicago : University of Chicago Press .

Ji , J. L. , Spronk , M. , Kulkarni , K. , Repov š , G. , Anticevic , A. , and Cole , M. W. (2019). Mapping the human brain ′ s cortical-subcortical functional network organization . *NeuroImage* , ***185*** , 35 – 57 . doi: 10.1016/j.neuroimage.2018.10.006

Jirsa , V. K. , and McIntosh , A. R. (eds.). (2007). *The Handbook of Brain Connectivity* . Berlin : Springer .

Johnson , K. (2004). Gold ′ s theorem and cognitive science . *Philosophy of Science* , ***71*** , 571 – 92 .

Johnson-Laird , P. N. (1988). *Computer and the Mind: An Introduction to Cognitive Science* . Cambridge, MA : Harvard University Press .

Jones , J. , and Roth , D. (2003). *Robot Programming: A Practical Guide to Behavior-Based Robotics* . New York : McGraw-Hill .

Kable , J. W. , and Glimcher , P. W. (2009). The neurobiology of decision: consensus and controversy . *Neuron* , ***63*** , 733 – 45 .

Kalat , J. W. (2010). *Introduction to Psychology* (9th ed.). Belmont, CA : Wadsworth Thomson Learning .

Kandel , E. R. , Schwarz , J. H. , and Jessell , T. M. (2012). *Principles of Neural Science* (5th ed.). New York : McGraw-Hill Medical .

Kanwisher , N. (2000). Domain specificity in face perception . *Nature Neuroscience* , ***3*** , 759 – 63 .

Kanwisher , N. , McDermott , J. , and Chun , M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for the perception of faces . *Journal of Neuroscience* , ***17*** , 4302 – 11 .

Kaplan , M. (1996). *Decision Theory as Philosophy* . Cambridge : Cambridge University Press .

Kelly , W. M. , Macrae , C. N. , Wyland , C. L. , Caglar , S. , Inati , S. , and Heatherton , T. F. (2002). Finding the self? An event-related fMRI study . *Journal of Cognitive Neuroscience* , ***14*** , 785 – 94 .

Kersten , D. , Mamassian , P. , and Yuille , A. (2004). Object perception as Bayesian inference . *Annual Review of Psychology* , ***55*** , 271 – 304 .

Kilner , J. M. , and Lemon , R. N. (2013). What we know currently about mirror neurons . *Current Biology* , ***23*** , R1057 – 62 .

Kiran , S. , and Lebel , K. (2007). Crosslinguistic semantic and translation priming in normal bilingual individuals and bilingual aphasia . *Clinical Linguistics and Phonetics* , ***4*** , 277 – 303 .

Knill , D. C. , and Whitman , R. (2008). *Perception as Bayesian Inference* . Cambridge : Cambridge University Press .

Koch , C. (2004). *The Quest for Consciousness: A Neurobiological Approach* . Englewood, CO : Roberts .

Koch , C. , Massimini , M. , Boly , M. , and Tononi , G. (2016). Neural correlates of consciousness: Progress and problems . *Nature Reviews Neuroscience* , ***17*** , 307 .

Koch , C. , and Tsuchiya , N. (2007). Attention and consciousness: Two distinct brain processes . *Trends in Cognitive Sciences* , ***11*** , 229 – 35 .

Koh , J.-S. , Yang , E. , Jung , G.-P. , Jung , S.-P. , Son , J. H. , Lee , S.-I. , … Cho , K.-J. (2015). Jumping on water: Surface tension – dominated jumping of water striders and robotic insects . *Science* , ***349*** , 517 – 21 .

Körding , K. P. , and Wolpert , D. M. (2004). Bayesian integration in sensorimotor learning . *Nature* , ***427*** , 244 .

Körding , K. P. , and Wolpert , D. M. (2006). Bayesian decision theory in sensorimotor control . *Trends in Cognitive Sciences* , ***10*** , 319 – 26 .

Kosslyn , S. M. (1973). Scanning visual images: Some structural implications . *Perception and Psychophysics* , ***14*** , 341 – 70 .

Kosslyn , S. M. , Thompson , W. L. , and Ganis , G. (2006). *The Case for Mental Imagery* . Oxford : Oxford University Press .

Kotseruba , I. , and Tsotsos , J. K. (2020). 40 years of cognitive architectures: core cognitive abilities and practical applications . *Artificial Intelligence Review* , ***53*** , 17 – 94 .

Kotz , S. A. (2001). Neurolinguistic evidence for bilingual language representation: A comparison of reaction times and event-related brain potentials . *Bilingualism: Language and Cognition* , ***4*** , 143 – 54 .

Kouider , S. , de Gardelle , V. , Sackur , J. , and Dupoux , E. (2010). How rich is consciousness? The partial awareness hypothesis . *Trends in Cognitive Sciences* , ***14*** , 301 – 7 .

Kouider , S. , and Dehaene , S. (2007). Levels of processing during non-conscious perception: A critical review of visual masking . *Philosophical Transactions of the Royal Society of London B* , ***362*** , 857 – 75 .

Kouider , S. , Dehaene , S. , Jobert , A. , and Le Bihan , D. (2007). Cerebral bases of subliminal and supraliminal priming during reading . *Cerebral Cortex* , ***17*** , 2019 – 29 .

Krichmar , J. L. (2018). Neurorobotics — a thriving community and a promising pathway toward intelligent cognitive robots . *Frontiers in Neurorobotics* , ***12*** . doi: 10.3389/fnbot.2018.00042

Kriegeskorte , N. , and Douglas , P. K. (2018). Cognitive computational neuroscience . *Nature Neuroscience* , ***21*** , 1148 – 60 . doi: 10.1038/s41593-018-0210-5

Krizhevsky , A. , Sutskever , I. , and Hinton , G. E. (2012). Image classification with deep convolutional neural networks. Paper presented at Advances in Neural Information Processing Systems (NIPS 2012).

Kurt , W. (2019). *Bayesian Statistics the Fun Way: Understanding statistics and probability with Star Wars, LEGO, and rubber ducks* . San Francisco : No Starch Press .

LaBar , K. S. , Gatenby , J. C. , Gore , J. C. , LeDoux , J. E. , and Phelps , E. A. (1998). Human amygdala activation during conditioned fear acquisition

and extinction: A mixed-trial fMRI study . *Neuron* , ***20*** , 937 – 45 .

Lake , B. , Salakhutdinov , R. , and Tenenbaum , J. B. (2016). Human level concept learning through probabilistic program induction . *Science* , ***350*** , 1332 – 8 .

Lake , B. M. , Ullman , T. D. , Tenenbaum , J. B. , and Gershman , S. J. (2017). Building machines that learn and think like people . *Behavioral and Brain Sciences* , ***40*** , e253 .

Lamme , V. A. F. (2003). Why visual attention and awareness are different . *Trends in Cognitive Sciences* , *7* , 12 – 8 .

Lamme , V. A. F. (2006). Towards a true neural stance on consciousness . *Trends in Cognitive Sciences* , ***10*** , 494 – 501 .

Langer , C. , and Ay , N. (2021). How morphological computation shapes integrated information in embodied agents . *Frontiers in Psychology* , ***12*** .

Lashley , K. S. (1951). The problem of serial order in behavior . In A. L. Jeffress (ed.), *Cerebral Mechanisms in Behavior: The Hixon Symposium* . New York : Wiley .

Lau , H. , and Rosenthal , D. (2011). Empirical support for higher-order theories of conscious awareness . *Trends in Cognitive Science* , ***15*** , 365 – 73 .

Laureys , S. (2005). The neural correlate of (un)awareness: Lessons from the vegetative state . *Trends in Cognitive Sciences* , ***9*** , 556 – 9 .

Lavie , N. (2005). Distracted and confused? Selective attention under load . *Trends in Cognitive Sciences* , ***9*** , 75 – 82 .

Lazarou , I. , Nikolopoulos , S. , Petrantonakis , P. C. , Kompatsiaris , I. , and Tsolaki , M. (2018). EEG-based brain – computer interfaces for communication and rehabilitation of people with motor impairment: A novel approach of the 21st century . *Frontiers in Human Neuroscience* , ***12*** (14).

Lazarus , R. S. (1991). *Emotion and Adaptation* . Oxford University Press .

Lebedev , M. , Opris , I. , and Casanova , M. F. (eds.). (2018). *Augmentation of Brain Function: Facts, Fiction and Controversy. Volume I: Brain-Machine Interfaces* . Frontiers Media SA . www.frontiersin.org/research-topics/1563/augmentation-of-brain-function-facts-fiction-and-controversy#overview

Lebiere , C. (2003). ACT . In L. Nadel (ed.), *Encyclopedia of Cognitive Science* . New York : Nature .

LeCun , Y. , Bengio , Y. , and Hinton , G. (2015). Deep learning . *Nature* , ***521*** , 436 .

LeDoux , J. E. (2000). Emotion Circuits in the Brain . *Annual Review of Neuroscience* , ***23*** , 155 – 84 .

LeDoux , J. E. (2012). Rethinking the emotional brain . *Neuron* , ***73*** (4), 653 – 76 .

Lerner , J. S. , Li , Y. , Valdesolo , P. , and Kassam , K. S. (2015). Emotion and decision making . *Annual Review of Psychology* , ***66*** , 799 – 823 .

Leslie , A. M. (1987). Pretense and representation: The origins of " theory of mind. " *Psychological Review* , ***94*** , 412 – 26 .

Leslie , A. M. , Friedman , O. , and German , T. P. (2004). Core mechanisms in " theory of mind. " *Trends in Cognitive Sciences* , ***8*** , 529 – 33 .

Leslie , A. M. , German , T. P. , and Polizzi , P. (2005). Belief – desire reasoning as a process of selection . *Cognitive Psychology* , ***50*** , 45 – 85 .

Leslie , A. M. , and Polizzi , P. (1998). Inhibitory processing in the false belief task: Two conjectures . *Developmental Science* , ***1*** , 247 – 53 .

Levenson , R. W. , Ekman , P. , and Friesen , W. V. (1990). Voluntary facial action generates emotion-specific autonomic nervous system activity . *Psychophysiology* , ***27*** , 363 – 84 . doi.10.1111/j.1469-8986.1990.tb02330.x

Levine , J. (1983). Materialism and qualia: The explanatory gap . *Pacific Philosophical Quarterly* , ***64*** , 354 – 61 .

Libby , T. , Moore , T. Y. , Chang-Siu , E. , Li , D. , Cohen , D. J. , Jusufi , A. , and Full , R. J. (2012). Tail-assisted pitch control in lizards, robots and dinosaurs . *Nature* , ***481*** , 181 – 4 .

Lidz , J. , Waxman , S. , and Freedman , J. (2003). What infants know about syntax but couldn ' t have learned: Experimental evidence for syntactic structure at 18 months . *Cognition* , ***89*** , B65 – 73 .

Lieto , A. (2021). *Cognitive Design for Artificial Minds* . Milton Park : Routledge .

Lillard , A. S. (2017). Why do the children (pretend) play? *Trends in Cognitive Sciences* , ***21*** , 826 – 34 . doi: 10.1016/j.tics.2017.08.001

Lillicrap , T. P. , Santoro , A. , Marris , L. , Akerman , C. J. , and Hinton , G. (2020). Backpropagation and the brain . *Nature Reviews Neuroscience* , ***21*** , 335 – 46 .

Logothetis , N. K. (2001). The underpinnings of the BOLD functional magnetic resonance imaging signal . *Journal of Neuroscience* , ***23*** , 3963 – 71 .

Logothetis , N. K. (2002). The neural basis of the blood-oxygen-level-dependent functional magnetic resonance imaging signal . *Philosophical Transactions of the Royal Society of London B* , ***357*** , 1003 – 37 .

Logothetis , N. K. (2008). What we can do and what we cannot do with fMRI . *Nature* , ***453*** , 869 – 78 .

Logothetis , N. K. , Pauls , J. , Augath , M. , Trinath , T. , and Oeltermann , A. (2001). Neurophysiological investigation of the fMRI signal . *Nature* , ***412*** , 150 – 7 .

Low , J. , and Perner , J. (2012). Implicit and explicit theory of mind: State of the art . *British Journal of Developmental Psychology* , ***30*** , 1 – 30 .

Luck , S. J. (2014). *An Introduction to the Event-Related Potential Technique* . Cambridge, MA : MIT Press .

Luck , S. J. , and Ford , M. A. (1998). On the role of selective attention in visual perception . *Proceedings of the National Academy of Sciences USA* , ***95*** , 825 – 30 .

Luck , S. J. , and Kappenman , E. S. (2011). *The Oxford Handbook of Event-Related Potential Components* . Oxford : Oxford University Press .

Ludlow , P. , Nagasawa , Y. , and Stoljar , D. (eds.). (2004). *There ′ s Something about Mary* . Cambridge, MA : MIT Press .

Luger , G. G. (2001). *Artificial intelligence: Structures and strategies for complex problem solving* (4th ed.) Boston, MA : Addison-Wesley .

Luo , Y. , and Baillargeon , R. (2010). Toward a mentalistic account of early psychological reasoning . *Current Directions in Psychological Science* , ***19*** , 301 – 7 .

Luria , A. R. (1970). The functional organization of the brain . *Scientific American* , ***222*** , 66 – 72 .

Ma , L. , and Lillard , A. S. (2017). The evolutionary significance of pretend play: Two-year-olds ′ interpretation of behavioral cues . *Learning & Behavior* , ***45*** , 441 – 8 . doi: 10.3758/s13420-017-0285-y

Macdonald , C. , and Macdonald , G. (1995). *Connectionism* . Oxford : Blackwell .

Machery , E. (2008). Massive modularity and the flexibility of human cognition . *Mind and Language* , ***23*** , 263 – 72 .

Machery , E. (2012). Dissociations in neuropsychology and cognitive neuroscience . *Philosophy of Science* , ***79*** , 490 – 518 .

Mack , A. , and Rock , I. (1998). *Inattentional Blindness* . Cambridge, MA : MIT Press .

Maeng , L. Y. , and Milad , M. R. (2017). Post-traumatic stress disorder: The relationship between the fear response and chronic stress . *Chronic Stress* , ***1*** , 2470547017713297. doi: 10.1177/2470547017713297

Manassi , M. , Sayim , B. , and Herzog , M. H. (2013). When crowding of crowding leads to uncrowding . *Journal of Vision* , ***13*** , 10. doi: 10.1167/13.13.10

Marcus , G. (2003). *The Algebraic Mind: Integrating Connectionism and Cognitive Science* . Cambridge, MA : MIT Press .

Marcus , G. , Ullman , M. , Pinker , S. , Hollander , M. , Rosen , T. J. , and Xu , F. (1992). *Overregularization in Language Acquisition* . Chicago : University of Chicago Press .

Mareschal , D. , and Johnson , S. P. (2002). Learning to perceive object unity: A connectionist account . *Developmental Science* , ***5*** , 151 – 85 .

Mareschal , D. , Plunkett , K. , and Harris , P. (1995). Developing object permanence: A connectionist model . In J. D. Moore and J. F. Lehman (eds.), *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* . Mahwah, NJ : Lawrence Erlbaum .

Margolis , E. , Samuels , R. , and Stich , S. (eds.). (2012). *The Oxford Handbook of Philosophy of Cognitive Science* . Oxford : Oxford University Press .

Marr , D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* . San Francisco : W. H. Freeman .

Marr , D. (2010). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* . 1982; repr., Cambridge, MA : MIT Press .

Marshall , J. C. , and Halligan , P. W. (1988). Blindsight and insight in visuospatial neglect . *Nature* , ***366*** , 766 – 7 .

Martens , S. , and Wyble , B. (2010). The attentional blink: Past, present, and future of a blind spot in perceptual awareness . *Neuroscience and Biobehavioral Reviews* , ***34*** , 947 – 57 .

Mashour , G. A. , Roelfsema , P. , Changeux , J.-P. , and Dehaene , S. (2020). Conscious processing and the global neuronal workspace hypothesis . *Neuron* , ***105*** , 776 – 98 . doi: 10.1016/j.neuron.2020.01.026

Matarić , M. (1997). Behavior-based control: Examples from navigation, learning, and group behavior . *Journal of Experimental and Theoretical Artificial Intelligence* , ***9*** , 323 – 36 .

Matarić , M. (1998). Behavior-based robotics as a tool for synthesis of artificial behavior and analysis of natural behavior . *Trends in Cognitive Sciences* , ***2*** , 82 – 7 .

Matarić, M. (2007). *The Robotics Primer* . Cambridge, MA : MIT Press .

Mathersul , D. , McDonald , S. , and Rushby , J. A. (2013). Understanding advanced theory of mind and empathy in high-functioning adults with autism spectrum disorder . *Journal of Clinical and Experimental Neuropsychology* , ***35*** , 655 – 68 .

McClelland , J. L. , Botvinick , M. M. , Noelle , D. C. , et al. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to cognition . *Trends in Cognitive Sciences* , ***14*** , 348 – 56 .

McClelland , J. L. , and Jenkins , E. (1991). Nature, nurture, and connectionism: Implications for connectionist models of development . In K. van Lehn (ed.), *Architectures for Intelligence: The 22nd (1988) Carnegie Symposium on Cognition* . Hillsdale, NJ : Lawrence Erlbaum .

McClelland , J. L. , and Patterson , K. (2002). Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences* , ***6*** , 465 – 72 .

McClelland , J. L , Rumelhart , D. E. , and the PDP Research Group (1986). *Parallel Distributed Processing: Explorations in the Microstructures of Cognition. Vol. 2: Psychological and Biological Models* . Cambridge, MA : MIT Press .

McCulloch , W. S. , and Pitts , W. H. (1943). A logical calculus of the ideas immanent in nervous activity . *Bulletin of Mathematical Biophysics* , ***5*** , 115 – 33 .

McDermott , J. H. (2009). The cocktail party problem . *Current Biology* , ***19*** , R1024 – 7 .

McLeod , P. , Plunkett , K. , and Rolls , E. T. (1998). *Introduction to the Connectionist Modelling of Cognitive Processes* . Oxford : Oxford University Press .

McSweeney , F. K. , and Murphy , E. S. (2014). *The Wiley-Blackwell Handbook of Classical and Operant Conditioning* . Chichester, UK : Wiley-Blackwell .

Medsker , L. R. , and Schulte , T. W. (2003). Expert systems . In L. Nadel (ed.), *Encyclopedia of Cognitive Science* (vol. 2). New York : Nature .

Melcher , D. , and Colby , C. L. (2008). Trans-saccadic perception . *Trends in Cognitive Sciences* , ***12*** , 466 – 73 .

Metzinger , T. (ed.). (2000). *Neural Correlates of Consciousness: Empirical and Conceptual Issues* . Cambridge, MA : MIT Press .

Michalski , R. S. , and Chilausky , R. L. (1980). Learning by being told and learning from examples: An experimental comparison of the two methods for knowledge acquisition in the context of developing an expert system for soybean disease diagnosis . *International Journal of Policy Analysis and Information Systems* , ***4*** , 125 – 61 .

Miller , B. , and Feil-Seifer , D. (2016). Embodiment, situatedness, and morphology for humanoid robots interacting with people . In A. Goswami and P. Vadakkepat (eds.), *Humanoid Robotics: A Reference* . Dordrecht, Netherlands : Springer .

Miller , G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information . *Psychological Review* , ***63*** , 81 – 97 .

Miller , G. A. (2003). The cognitive revolution: A historical perspective . *Trends in Cognitive Sciences* , ***7*** , 141 – 4 .

Millican , P. , and Clark , A. (1996). *The Legacy of Alan Turing: Machines and Thought* . London : Clarendon Press .

Milner , A. D. (2012). Is visual processing in the dorsal stream accessible to consciousness? *Proceedings of the Royal Society B* , ***279*** , 2289 – 98 .

Milner , A. D. , and Goodale , M. A. (1998). *The Visual Brain in Action* (Precis) . *Psyche* , ***4*** .

Milner , A. D. , and Goodale , M. A. (2006). *The Visual Brain in Action* (2nd ed.). Oxford : Oxford University Press .

Milner , A. D. , and Goodale , M. A. (2008). Two visual systems reviewed . *Neuropsychologia* , ***46*** , 774 – 85 .

Minati , L. , Varotto , G. , D ′ Incerti , L. , Panzica , F. , and Chan , D. (2013). From brain topography to brain topology: Relevance of graph theory to functional neuroscience . *Neuroreport* , ***24*** , 536 – 43 .

Minsky , M. , and Papert , S. (1969). *Perceptrons* . Cambridge, MA : MIT Press .

Mishkin , M. L. , Ungerleider , G. , and Macko , K. A. (1983). Object vision and spatial vision: Two cortical pathways . *Trends in Neurosciences* , *6* , 414 – 7. Reprinted in Bechtel, W., Mandik, P., Mundale, J., and Stufflebeam, R. (eds.). (2001). *Philosophy and the Neurosciences: A Reader* . Oxford : Blackwell .

Mitchell , J. P. , Banaji , M. R. , and Macrae , C. N. (2005). The link between social cognition and self-referential thought in the medial prefrontal cortex . *Journal of Cognitive Neuroscience* , ***17*** , 1306 – 15 .

Mitchell , T. M. (1997). *Machine Learning* . Boston : McGraw-Hill .

Mobbs , D. , Yu , R. , Rowe , J. B. , Eich , H. , FeldmanHall , O. , and Dalgleish , T. (2010). Neural activity associated with monitoring the oscillating threat value of a tarantula . *Proceedings of the National Academy of Sciences* , ***107*** , 20582 – 6 . doi: 10.1073/pnas.1009076107

Mohan , A. , Roberto , A. J. , Mohan , A. , Lorenzo , A. , Jones , K. , Carney , M. J. , … Lapidus , K. A. (2016). The significance of the default mode network (DMN) in neurological and neuropsychiatric disorders: A review . *Yale Journal of Biology and Medicine* , ***89*** , 49 – 57 .

Molenberghs , P. , Sale , M. V. , and Mattingley , J. B. (2012). Is there a critical lesion site for unilateral spatial neglect? A meta-analysis using activation likelihood estimation . *Frontiers in Human Neuroscience* , ***6*** , 1 – 10 .

Montague , P. , Dayan , P. , and Sejnowski , T. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning . *Journal of Neuroscience* , ***16*** , 1936 – 47 .

Moors , A. , Ellsworth , P. C. , Scherer , K. R. , and Frijda , N. H. (2013). Appraisal theories of emotion: State of the art and future development . *Emotion Review* , ***5*** , 119 – 24 .

Moors , A. , and Fischer , M. (2019). Demystifying the role of emotion in behaviour: toward a goal-directed account . *Cognition and Emotion* , ***33*** , 94 – 100 .

Morse , S. J. , and Roskies , A. L. (2013). *A Primer on Criminal Law and Neuroscience* . Oxford : Oxford University Press .

Mountcastle , V. B. , Lynch , J. C. , Georgopoulos , A. , Sakata , H. , and Acuna , C. (1975). Posterior parietal association cortex of the monkey: Command functions for operations within extrapersonal space . *Journal of Neurophysiology* , ***38*** , 871 – 908 .

Mukamel , R. , Gelbard , H. , Arieli , A. , Hasson , U. , Fried , I. , and Malach , R. (2005). Coupling between neuronal firing, field potentials, and fMRI in human auditory cortex . *Science* , ***309*** , 951 – 4 .

Munakata , Y. (2001). Graded representations in behavioral dissociations . *Trends in Cognitive Sciences* , ***5*** , 309 – 15 .

Munakata , Y. , and McClelland , J. L. (2003). Connectionist models of development . *Developmental Science* , ***6*** , 413 – 29 .

Munakata , Y. , McClelland , J. L. , Johnson , M. H. , and Siegler , R. S. (1997). Rethinking infant knowledge: Toward an adaptive process account of successes and failures in object permanence tasks . *Psychological Review* , ***104*** , 686 – 713 .

Mundy , P. (2018). A review of joint attention and social-cognitive brain systems in typical development and autism spectrum disorder . *European Journal of Neuroscience* , ***47*** , 497 – 514 . doi: 10.1111/ejn.13720

Nadeau , S. E. (2021). Basal ganglia and thalamic contributions to language function: Insights from a parallel distributed processing perspective . *Neuropsychology Review* , ***31*** , 495 – 515 .

Nadel , L. (ed.). (2005). *Encyclopedia of Cognitive Science* . Chichester, UK : Wiley .

Nagel , T. (1974). What is it like to be a bat? *Philosophical Review* , ***83*** , 435 – 50 . doi: 10.2307/2183914

Needham , A. , and Libertus , K. (2011). Embodiment in early development . *WIREs Cognitive Science* , ***2*** , 117 – 23 .

Newmeyer , F. J. (1986). *Linguistic Theory in America* . London : Academic Press .

Nichols , S. , Stich , S. , Leslie , A. , and Klein , D. (1996). Varieties of off-line simulation . In P. Carruthers and P. K. Smith (eds.), *Theories of Theory of Mind* . Cambridge : Cambridge University Press .

Nicholson , T. , Williams , D. M. , Grainger , C. , Lind , S. E. , and Carruthers , P. (2019). Relationships between implicit and explicit uncertainty monitoring and mindreading: Evidence from autism spectrum disorder . *Consciousness and Cognition* , ***70*** , 11 – 24 . doi: 10.1016/j.concog.2019.01.013

Nicholson , T. , Williams , D. M. , Lind , S. E. , Grainger , C. , and Carruthers , P. (2021). Linking metacognition and mindreading: Evidence from autism and dual-task investigations . *Journal of Experimental Psychology: General* , ***150*** , 206 – 20 . doi: 10.1037/xge0000878

Nida-Rümelin , M. , and O Conaill , D. (2021). Qualia: The knowledge argument . In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab* , Stanford University . https://plato.stanford.edu/archives/sum2021/entries/qualia-knowledge/

Nilsson , N. J. (1984). *Shakey the robot [Technical Note 323]* . Menlo Park, CA : SRI International .

Norman , D. A. , and Shallice , T. (1980). Attention to action: Willed and automatic control of behaviour. Reprinted in Gazzaniga, M. (ed.). (2000). *Cognitive Neuroscience: A Reader* . Oxford : Blackwell .

Oakes , L. M. (2010). Using habituation of looking time to assess mental processes in infancy . *Journal of Cognition and Development* , ***11*** , 255 – 68 .

O ′ Grady , W. , Archibald , J. , Aronoff , M. , and Rees-Miller , J. (2010). *Contemporary Linguistics: An Introduction* (6th ed.). Boston : Bedford/St. Martin ′ s .

Onishi , K. H. , and Baillargeon , R. (2005). Do 15-month-old infants understand false beliefs? *Science* , ***308*** , 255 – 8 .

Op de Beeck , H. , and Nakatani , C. (2019). *Introduction to Human Neuroimaging* . Cambridge : Cambridge University Press .

Orban , G. A. , Van Essen , D. , and Vanduffel , W. (2004). Comparative mapping of higher visual areas in monkeys and humans . *Trends in Cognitive Sciences* , ***8*** , 315 – 24 .

O ′ Reilly , R. C. , and Munakata , Y. (2000). *Computational Explorations in Computational Neuroscience: Understanding the Mind by Simulating the Brain* . Cambridge, MA : MIT Press .

Owen , A. M. , Coleman , M. R. , Boly , M. , Davis , M. H. , Laureys , S. , and Pickard , J. D. (2006). Detecting awareness in the vegetative state . *Science* , ***313*** , 1402.

Page , M. (2000). Connectionist modeling in psychology: A localist manifesto . *Behavioral and Brain Sciences* , ***23*** , 443 – 67 .

Pakskepp , J. (2004). *Affective Neuroscience: The Foundations of Human and Animal Emotions* . Oxford : Oxford University Press .

Palmiero , M. , Piccardi , L. , Giancola , M. , Nori , R. , D ′ Amico , S. , and Olivetti Belardinelli , M. (2019). The format of mental imagery: from a critical review to an integrated embodied representation approach . *Cognitive Processing* , ***20*** , 277 – 89 . doi: 10.1007/s10339-019-00908-z

Panksepp , J. (2004). *Affective Neuroscience: The Foundations of Human and Animal Emotions* . Oxford : Oxford University Press .

Passingham , R. (2009). How good is the macaque monkey model of the human brain? *Current Opinion in Neurobiology* , ***19*** , 6 – 11 .

Pearl , L. , and Goldwater , S. (2016). Statistical learning, inductive bias, and Bayesian inference in language acquisition . In J. Lidz , W. Snyder , and C. Pater (eds.), *The Oxford Handbook of Developmental Linguistics* . Oxford : Oxford University Press .

Pearson , J. , and Kosslyn , S. M. (2015). The heterogeneity of mental representation: Ending the imagery debate . *Proceedings of the National Academy of Sciences USA* , ***112*** , 10089 – 92 .

Pelucchi , B. , Hay , J. F. , and Saffran , J. R. (2009). Statistical learning in a natural language by 8-month-old infants . *Child Development* , ***80*** , 674 – 85 .

Perner , J. (1993). *Understanding the Representational Mind* . Cambridge, MA : MIT Press .

Perner , J. , and Leekam , S. (2008). The curious incident of the photo that was accused of being false: Issues of domain specificity in development, autism, and brain imaging . *Quarterly Journal of Experimental Psychology* , ***61*** , 76 – 89 .

Perner , J. , and Roessler , J. (2012). From infants ′ to children ′ s appreciation of belief . *Trends in Cognitive Sciences* , ***16*** , 519 – 25 .

Perner , J. , and Ruffman , T. (2005). Infants ′ insight into the mind: How deep? *Science* , ***308*** , 214 – 6 .

Perone , S. , and Simmering , V. R. (2017). Applications of dynamic systems theory to cognition and development: New frontiers . *Advances in Child Development and Behavior* , ***52*** , 43 – 80 .

Peru , A. , Moro , V. , Avesani , R. , and Aglioti , S. (1996). Overt and covert processing of left-side information in unilateral neglect investigated with chimeric drawings . *Journal of Clinical and Experimental Neuropsychology* , ***18*** , 621 – 30 .

Petersen , S. E. , and Fiez , J. A. (2001). The processing of single words studied with positron emission tomography . In W. Bechtel , P. Mandik , J. Mundale , and R. S. Stufflebeam (eds.), *Philosophy and the Neurosciences: A Reader* . Malden, MA : Blackwell .

Petersen , S. E. , Fox , P. T. , Posner , M. I. , and Mintun , M. (1988). Positron emission tomographic studies of the cortical anatomy of single-word processing . *Nature* , ***331*** , 585 – 9 .

Petzold , C. (2008). *The Annotated Turing: A Guided Tour through Alan Turing ' s Historic Paper on Computability and the Turing Machine* . Indianapolis, IN : Wiley .

Pfeifer , R. , Iida , F. , and Gómez , G. (2006). Morphological computation for adaptive behavior and cognition . *International Congress Series* , ***1291*** , 22 – 9 .

Phelps , E. A. (2006). Emotion and cognition: Insights from studies of the human amygdala . *Annual Review of Psychology* , ***57*** , 27 – 53 .

Phillips , I. (2021). Blindsight is qualitatively degraded conscious vision . *Psychological Review* , ***128*** , 558 – 84 . doi: 10.1037/rev0000254

Phillips , J. , Buckwalter , W. , Cushman , F. , Friedman , O. , Martin , A. , Turri , J. , … Knobe , J. (2021). Knowledge before belief . *Behavioral and*

Brain Sciences , ***44*** , e140 . doi: 10.1017/S0140525X20000618

Phillips , M. L. , Young , A. W. , Senior , C. , Brammer , M. , Andrew , C. , Calder , A. J. , … David , A. S. (1997). A specific neural substrate for perceiving facial expressions of disgust . *Nature* , ***389*** , 495 – 8 .

Piaget , J. (1954). *The Construction of Reality in the Child* . New York : Basic Books .

Piccinini , G. (2004). The first computational theory of mind and brain: A close look at McCulloch and Pitts ′ “ Logical Calculus of the Ideas Immanent in Nervous Activity. ″ *Synthese* , ***141*** , 175 – 215 .

Piccinini , G. , and Craver , C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches . *Synthese* , ***183*** , 283 – 311 .

Pietraszewski , D. , and Wertz , A. E. (2021). Why evolutionary psychology should abandon modularity. *Perspectives on Psychological Science* . Advance online publication. doi: 10.1177/1745691621997113

Pillow , J. , and Sahani , M. (2019). Editorial overview: Machine learning, big data, and neuroscience . *Current Opinion in Neurobiology* , ***55*** , iii – iv . doi: 10.1016/j.conb.2019.05.002

Pinker , S. (1994). *The Language Instinct: How the Mind Creates Language* . New York : Harper .

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

Pinker , S. (2005). So how does the mind work? *Mind and Language* , ***20*** , 1 – 24 .

Pinker , S. , and Prince , A. (1988 a). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition . *Cognition* , ***28*** , 73 – 193 .

Pinker , S. , and Prince , A. (1988 b). Rules and connections in human language . In R. Morris (ed.), *Parallel Distributed Processing* . Oxford : Oxford University Press .

Pinker , S. , and Ullman , M. T. (2002). The past and future of the past tense . *Trends in Cognitive Sciences* , ***6*** , 456 – 63 .

Platt , M. L. , and Glimcher , P. W. (1999). Neural correlates of decision variables in parietal cortex . *Nature* , ***400*** , 233 – 8 .

Plaut , D. C. , Banich , M. T. , and Mack , M. (2003). Connectionist modeling of language: Examples and implications . In M. T. Banich and M. Mack (eds.), *Mind, Brain, and Language: Multidisciplinary Perspectives* . Mahwah, NJ : Lawrence Erlbaum .

Plaut , D. C. , and McClelland , J. L. (2010). Locating object knowledge in the brain: Comment on Bowers ′ s (2009) attempt to revive the grandmother cell hypothesis . *Psychological Review* , ***117*** , 284 – 90 .

Plotnik , R. , and Kouyoumdjian , H. (2010). *Introduction to Psychology* (9th ed.). Belmont, CA : Wadsworth Thomson Learning .

Plunkett , K. , and Elman , J. L. (1997). *Exercises in Rethinking Innateness: A Handbook for Connectionist Simulations* . Cambridge, MA : MIT Press .

Plunkett , K. , and Marchman , V. (1993). From rote learning to system building: Acquiring verb morphology in children and connectionist nets . *Cognition* , ***48*** , 21 – 69 .

Poldrack , R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences* , ***10*** , 59 – 63 .

Poldrack , R. A. (2018). *The New Mind Readers: What Neuroimaging Can and Cannot Reveal about Our Thoughts* . Princeton, NJ : Princeton University Press .

Poldrack , R. A. , Baker , C. I. , Durnez , J. , Gorgolewski , K. J. , Matthews , P. M. , Munafò , M. R. , … Yarkoni , T. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research . *Nature Reviews Neuroscience* , ***18*** , 115 – 26 .

Poldrack , R. A. , Mumford , J. A. , and Nichols , T. E. (2011). *Handbook of Functional MRI Data Analysis* . Cambridge : Cambridge University Press .

Poole , D. L. , and Mackworth , A. K. (2010). *Artificial Intelligence: Foundations of Computational Agents* . Cambridge : Cambridge University Press .

Pöppel , E. , Frost , D. , and Held , R. (1973). Residual visual function after brain wounds involving the central visual pathways in man . *Nature* ,

243 , 295 – 6 .

Port , R. F. , and Van Gelder , T. (1995). *Mind as Motion: Explorations in the Dynamics of Cognition* . Cambridge, MA : MIT Press .

Posner , J. , Russell , J. A. , and Peterson , B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology . *Developmental Psychopathology* , ***17*** , 715 – 34 .

Posner , M. I. (1980). Orienting of attention . *Quarterly Journal of Experimental Psychology* , ***32*** , 3 – 25 .

Posner , M. I. (1989). *Foundations of Cognitive Science* . Cambridge, MA : MIT Press .

Posner , M. I. (ed.). (2004). *The Cognitive Neuroscience of Attention* . New York : Guilford .

Posner , M. I. (2017). Attentional mechanisms. Reference Module in Neuroscience and Biobehavioral Psychology [online resource]. Elsevier.

Posner , M. I. , and Raichle , M. E. (1994). *Images of Mind* . New York : Scientific American Library .

Poulin-Dubois , D. , Brooker , I. , and Chow , V. (2009). The developmental origins of naive psychology in infancy . *Advances in Child Development and Behavior* , ***37*** , 55 – 104 .

Preston , J. M. , and Bishop , J. M. (2002). *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence* . Oxford : Oxford

University Press .

Prinz , J. (2012). *The Conscious Brain* . New York : Oxford University Press .

Pullum Geoffrey , K. , and Scholz Barbara , C. (2002). Empirical assessment of stimulus poverty arguments . *The Linguistic Review* , ***18*** , 9 – 50 .

Purves , D. , Augustine , G. J. , Fitzpatrick , D. , Hall , W. C. , Anthony-Samuel , L. , and White , L. E. (2011). *Neuroscience* (5th ed.). Sunderland, MA : Sinauer Associates .

Pylyshyn , Z. (1980). Computation and cognition: Issues in the foundations of cognitive science . *Behavioral and Brain Sciences* , ***3*** , 111 – 69 .

Pylyshyn , Z. (1984). *Computation and Cognition: Toward a Foundation for Cognitive Science* . Cambridge, MA : MIT Press .

Pylyshyn , Z. W. (2002). Mental imagery: In search of a theory . *Behavioral and Brain Sciences* , ***25*** , 157 – 82 . doi: 10.1017/S0140525X02000043

Quilty-Dunn , J. (2020). Perceptual pluralism . *Noûs* , ***54*** , 807 – 38 . doi: 10.1111/nous.12285

Quinlan , P. T. , van der Maas , H. L. J. , Jansen , B. R. J. , Booij , O. , and Rendell , M. (2007). Re-thinking stages of cognitive development: An appraisal of connectionist models of the balance scale task . *Cognition* , ***103*** , 413 – 59 .

Raffone , A. , Srinivasan , N. , and van Leeuwen , C. (2014). The interplay of attention and consciousness in visual search, attentional blink and working memory consolidation . *Philosophical Transactions of the Royal Society B* , ***369*** , 20130215.

Raichle , M. E. , and Mintun , M. A. (2006). Brain work and brain imaging . *Annual Review of Neuroscience* , ***29*** , 449 – 76 .

Rajalingham , R. , Issa , E. B. , Bashivan , P. , Kar , K. , Schmidt , K. , and DiCarlo , J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks . *Journal of Neuroscience* , ***38*** , 7255 – 7269 . doi: 10.1523/JNEUROSCI.0388-18.2018

Rakoczy , H. (2012). Do infants have a theory of mind? *British Journal of Development Psychology* , ***30*** , 59 – 74 .

Ramnani , N. , Behrens , T. E. J. , Penny , W. , and Matthews , P. M. (2004). New approaches for exploring functional and anatomical connectivity in the human brain . *Biological Psychiatry* , ***56*** , 613 – 9 .

Ramsey , W. , Stich , S. P. , and Rumelhart , D. E. (1991). *Philosophy and Connectionist Theory* . Hillsdale, NJ : Lawrence Erlbaum .

Rees , G. , Friston , K. , and Koch , C. (2000). A direct quantitative relationship between the functional properties of human and macaque V5 . *Nature Neuroscience* , ***3*** , 716 – 23 .

Regier , T. , and Gahl , S. (2004). Learning the unlearnable: The role of missing evidence . *Cognition* , ***93*** , 147 – 55 .

Riley , M. A. , and Holden , J. G. (2012). Dynamics of cognition. *WIREs Cognitive Science* , ***3*** , 593 – 606 .

Rizzolatti , G. , and Fogassi , L. (2014). The mirror mechanism: Recent findings and perspectives . *Philosophical Transactions of the Royal Society of London B* , ***369*** , 20130420.

Rizzolatti , G. , Fogassi , L. , and Gallese , V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action . *Nature Reviews Neuroscience* , ***2*** , 661 – 70 .

Rizzolatti , G. , Fogassi , L. , and Gallese , V. (2006). Mirrors of the mind . *Scientific American* , ***295*** , 54 – 61 .

Rizzolatti , G. , and Sinigaglia , C. (2008). *Mirrors in the Brain: How Our Minds Share Actions and Emotions* . Trans. F. Anderson . Oxford : Oxford University Press .

Rizzolatti , G. , and Sinigaglia , C. (2010). The functional role of the parieto-frontal mirror circuit: Interpretations and misinterpretations . *Nature Reviews Neuroscience* , ***11*** , 264 – 74 .

Robbins , R. , and Aydede , M. (eds.). (2008). *The Cambridge Handbook of Situated Cognition* . Cambridge : Cambridge University Press .

Robinson , D. L. , Goldberg , M. E. , and Stanton , G. B. (1978). Parietal association cortex in the primate: Sensory mechanisms and behavioral modulations . *Journal of Neurophysiology* , ***41*** , 910 – 32 .

Roeyers , H. , and Demurie , E. (2010). How impaired is mind-reading in high-functioning adolescents and adults with autism? *European Journal of Developmental Psychology* , ***7*** , 123 – 34 .

Rogers , R. (1971). *Mathematical Logic and Formalized Theories* . Amsterdam : North-Holland .

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

Rohde , D. , and Plaut , D. C. (1999). Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition* , ***72*** , 67 – 109 .

Rollins , M. (1989). *Mental Imagery: The Limits of Cognitive Science* . Cambridge, MA : MIT Press .

Rolls , E. T. (2021). Learning invariant object and spatial view representations in the brain using slow unsupervised learning . *Frontiers in Computational Neuroscience* , ***15*** , 686239 .

Rolls , E. T. , and Milward ., T. (2000). A model of invariant object recognition in the visual system: Learning rules, activation functions, lateral inhibition, and information-based performance measures . *Neural Computation* , ***12*** , 2547 – 72 .

Rosenblatt , F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain . *Psychological Review* , ***65*** , 386 – 408 .

Rosenthal , D. M. (1986). Two concepts of consciousness . *Philosophical Studies* , ***49*** , 329 – 59 .

Rosenthal , D. M. (1997). A theory of consciousness . In N. Block , O. J. Flanagan , and G. Guzeldere (eds.), *The Nature of Consciousness* . Cambridge, MA : MIT Press .

Rossetti , Y. , Pisella , L. , and McIntosh , R. D. (2017). Rise and fall of the two visual systems theory . *Annals of Physical and Rehabilitation Medicine* , ***60*** , 130 – 40 .

Rowe , J. B. , and Frackowiak , R. S. J. (2003). Neuroimaging . In L. Nadel (ed.), *Encyclopedia of Cognitive Science* . New York : Nature .

Rumelhart , D. E. (1989). The architecture of mind: A connectionist approach . In M. I. Posner (ed.), *Foundations of Cognitive Science* . Cambridge, MA : MIT Press . Reprinted in Haugeland, J. (ed.). (1997). *Mind Design II: Philosophy, Psychology, Artificial Intelligence* . Cambridge, MA: MIT Press.

Rumelhart , D. E. , and McClelland , J. L. (1986). On learning the past tenses of English verbs . In J. L. McClelland , D. E. Rumelhart , and the PDP Research Group (eds.), *Parallel Distributed Processing: Explorations in the Microstructures of Cognition: Vol. 2. Psychological and Biological Models* . Cambridge, MA : MIT Press .

Rumelhart , D. E. , McClelland , J. L. , and the PDP Research Group. (eds.). (1986). *Parallel Distributed Processing: Explorations in the Microstructures of Cognition: Vol. 1. Foundations* . Cambridge, MA : MIT Press .

Russell , S. J. , and Norvig , P. (2003). *Artificial Intelligence: A Modern Approach* (2nd ed.). Upper Saddle River, NJ : Prentice Hall .

Russell , S. J. , and Norvig , P. (2009). *Artificial Intelligence: A Modern Approach* (3rd ed.). New Delhi : Prentice Hall .

Ryczko , D. , Simon , A. , and Ijspeert , A. J. (2020). Walking with salamanders: From molecules to biorobotics . *Trends in Neurosciences* , ***43*** , 916 – 30 .

Saffran , J. R. , Aslin , R. N. , and Newport , E. L. (1996). Statistical learning by 8-month-old infants . *Science* , **274** , 1926 – 8 .

Samson , D. , Apperly , I. A. , Chiavarino , C. , and Humphreys , G. W. (2004). Left temporoparietal junction is necessary for representing someone else ′ s belief . *Nature Neuroscience* , *7* , 499 – 500 .

Samson , D. , Apperly , I. A. , Kathirgamanathan , U. , and Humphreys , G. W. (2005). Seeing it my way: A case of a selective deficit in inhibiting self-perspective. *Brain: A Journal of Neurology* , ***128*** , 1102 – 11 .

Samuelson , L. K. , Jenkins , G. W. , and Spencer , J. P. (2015). Grounding cognitive-level processes in behavior: The view from dynamic systems theory . *Topics in Cognitive Science* , *7* , 191 – 205 .

Savopoulos , P. , and Lindell , A. K. (2018). Born criminal? Differences in structural, functional and behavioural lateralization between criminals and noncriminals . *Laterality* , ***23*** , 738 – 60 .

Saxe , R. (2009). Theory of mind (neural basis) . In W. Banks (ed.), *Encyclopedia of Consciousness* . Cambridge, MA : MIT Press .

Saxe , R. , Carey , S. , and Kanwisher , N. (2004). Understanding other minds: Linking developmental psychology and functional neuroimaging . *Annual Review of Psychology* , ***55*** , 87 – 124 .

Saxe , R. , and Kanwisher , N. (2005). People thinking about thinking people: The role of the temporo-parietal junction in " theory of mind. " In J. T. Cacioppo and G. G. Berntson (eds.), *Social Neuroscience: Key Readings* . New York : Psychology Press .

Schenk , T. , and McIntosh , R. D. (2010). Do we have independent visual streams for perception and action? *Cognitive Neuroscience* , ***1*** , 52 – 78 .

Schlatter , M. , and Aizawa , K. (2008). Walter Pitts and " A logical calculus. " *Synthese* , ***162*** , 235 – 50 .

Schneider , S. (2011). *The Language of Thought: A New Philosophical Direction* . Cambridge, MA : MIT Press .

Schneider , S. , and Katz , M. (2012). Rethinking the language of thought . *WIREs Cognitive Science* , ***3*** , 153 – 62 .

Schomer , D. L. , and Lopes da Silva , F. H. (2018). *Niedermeyer ' s Encephalography: Basic Principles, Clinical Applications, and Related Fields* . New York : Oxford University Press .

Schoonbaert , S. , Duyck , W. , Brysbaert , M. , and Hartsuiker , R. J. (2009). Semantic and translation priming from a first language to a second and

back: Making sense of the findings . *Memory & Cognition* , ***37*** , 569 – 86 .

Schultz , W. , Dayan , P. , and Montague , P. R. (1997). A neural substrate of prediction and reward . *Science* , ***275*** , 1593 – 9 . doi: 10.1126/science.275.5306.1593

Schurz , M. , and Perner , J. (2015). An evaluation of neurocognitive models of theory of mind . *Frontiers in Psychology* , ***6*** , 1610 .

Schuwerk , T. , Priewasser , B. , Sodian , B. , and Perner , J. (2018). The robustness and generalizability of findings on spontaneous false belief sensitivity: A replication attempt . *Royal Society Open Science* , ***5*** , 172273 . doi: 10.1098/rsos.172273

Schyns , P. G. , Gosselin , F. , and Smith , M. L. (2008). Information processing algorithms in the brain . *Trends in Cognitive Sciences* , ***13*** , 20 – 6 .

Searle , J. (1980). Minds, brains, and programs . *Behavioral and Brain Sciences* , ***3*** , 417 – 57 .

Searle , J. (2004). *Mind: A Brief Introduction* . New York : Oxford University Press .

Sehlmeyer , C. , Schöning , S. , Zwitserlood , P. , Pfleiderer , B. , Kircher , T. , Arolt , V. , … Konrad , C. (2009). Human fear conditioning and extinction in neuroimaging: a systematic review . *PloS One* , ***4*** , e5865 .

Shadmehr , R. , and Krakauer , J. W. (2008). A computational neuroanatomy for motor control . *Experimental Brain Research* , **1** ***85*** , 359 – 81 .

Shanahan , M. P. (2003). The frame problem . In L. Nadel (ed.), *Encyclopedia of Cognitive Science* . New York : Nature .

Shannon , C. E. (1948). A mathematical theory of communication . *Bell System Technical Journal* , ***27*** , 379 – 423, 623 – 56 .

Shapiro , L. (2007). The embodied cognition research programme . *Philosophy Compass* , ***2*** , 338 – 46 .

Shapiro , L. (2011). *Embodied Cognition* . New York : Routledge .

Shechner , T. , Hong , M. , Britton , J. C. , Pine , D. S. , and Fox , N. A. (2014). Fear conditioning and extinction across development: evidence from human studies and animal models . *Biological Psychology* , ***100*** , 1 – 12 .

Shepard , R. N. , and Metzler , J. (1971). Mental rotation of three-dimensional objects . *Science* , ***171*** , 701 – 3 .

Shepherd , G. (1994). *Neurobiology* (3rd ed.). New York : Oxford University Press .

Siegelmann , H. , and Sontag , E. (1991). Turing computability with neural nets . *Applied Mathematics Letters* , ***4*** , 77 – 80 .

Silver , D. , Huang , A. , Maddison , C. J. , Guez , A. , Sifre , L. , van den Driessche , G. , … Hassabis , D. (2016). Mastering the game of Go with deep neural networks and tree search . *Nature* , ***529*** , 484 .

Silver , D. , Schrittwieser , J. , Simonyan , K. , Antonoglou , I. , Huang , A. , Guez , A. , … Hassabis , D. (2017). Mastering the game of Go without human knowledge . *Nature* , ***550*** , 354 .

Simon , H. (1967). Motivational and emotional controls of cognition . *Psychological Review* , ***74*** , 29 – 39 .

Simonet , C. , and McNally , L. (2021). Kin selection explains the evolution of cooperation in the gut microbiota . *Proceedings of the National Academy of Sciences* , ***118*** , e2016046118 .

Simons , D. , and Chabris , C. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events . *Perception* , ***28*** , 1059 – 74 .

Simons , D. , and Rensink , R. A. (2005). Change blindness: Past, present, and future . *Trends in Cognitive Sciences* , ***9*** , 16 – 20 .

Singer , W. (1999). Neuronal synchrony: A versatile code for the definition of relations? *Neuron* , ***24*** , 49 – 65 .

Skerry , A. E. , and Saxe , R. (2015). Neural representations of emotion are organized around abstract event features . *Current Biology* , ***25*** , 1945 – 54 .

Skyrms , B. (1986). *Choices and Chance* . Belmont, CA : Wadsworth .

Smith , L. , and Thelen , E. (2003). Development as a dynamical system . *Trends in Cognitive Sciences* , *7* , 343 – 8 .

Spelke , E. S. (1988). The origins of physical knowledge . In L. Weiskrantz (ed.), *Thought without Language* . Oxford : Oxford University

Press .

Spelke , E. S. , Gutheil , G. , Van de Walle , G. , Kosslyn , S. M. , and Osherson , D. N. (1995). The development of object perception . In S. M. Kosslyn and D. N. Osherson (eds.), *An Invitation to Cognitive Science: Vol. 2. Visual Cognition* (2nd ed.). Cambridge, MA : MIT Press .

Spelke , E. S. , and Kinzler , K. D. (2007). Core knowledge . *Developmental Science* , ***10*** , 89 – 96 .

Spelke , E. S. , and Van de Walle , G. (1993). Perceiving and reasoning about objects: Insights from infants . In N. Eilan , R. McCarthy , and B. Brewer (eds.), *Spatial Representation* . Oxford : Blackwell .

Spencer , J. P. , Austin , A. , and Schutte , A. R. (2012). Contributions of dynamic systems theory to cognitive development . *Cognitive Development* , ***27*** , 401 – 18 .

Spencer , J. P. , Perone , S. , and Buss , A. T. (2011). Twenty years and going strong: A dynamic systems revolution in motor and cognitive development . *Child Development Perspectives* , ***5*** , 260 – 6 .

Spencer , J. P. , Thomas , M. S. C. , and McClelland , J. L. (2009). *Toward a Unified Theory of Development: Connectionism and Dynamic Systems Theory Reconsidered* . New York : Oxford University Press .

Sperling , G. (1960). The information available in brief visual presentations . *Psychological Monographs* , ***74*** , 1 – 29 .

Spivey , M. (2007). *The Continuity of Mind* . New York : Oxford University Press .

Spunt , R. P. , and Adolphs , R. (2017). A new look at domain specificity: insights from social neuroscience . *Nature Reviews Neuroscience* , ***18*** , 559 – 67 .

Stein , J. F. , and Stoodley , C. S. (2006). *Neuroscience: An Introduction* . Oxford : Oxford University Press .

Sterelny , K. (1990). *The Representational Theory of Mind* . Oxford : Blackwell .

Sullivan , J. A. (2009). The multiplicity of experimental protocols: A challenge to reductionist and non-reductionist models of the unity of neuroscience . *Synthese* , ***167*** , 511 – 39 .

Sun , R. (ed.). (2008). *The Cambridge Handbook of Computational Psychology* . Cambridge : Cambridge University Press .

Sutton , R. S. , and Barto , A. G. (1998). *Reinforcement Learning: An Introduction* . Cambridge MA : MIT Press .

Sutton , R. S. , and Barto , A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). Cambridge MA : MIT Press .

Taddeo , M. , and Floridi , L. (2005). Solving the symbol grounding problem: a critical review of fifteen years of research . *Journal of Experimental & Theoretical Artificial Intelligence* , ***17*** , 419 – 45 .

Tamir , D. I. , and Mitchell , J. P. (2010). Neural correlates of anchoring-and-adjustment during mentalizing . *Proceedings of the National Academy of Sciences USA* , ***107*** , 10827 – 32 .

Thelen , E. , Schöner , G. , Scheier , C. , and Smith , L. B. (2001). The dynamics of embodiment: A field theory of infant perseverative reaching . *Behavioral and Brain Sciences* , ***24*** , 1 – 86 .

Thelen , E. , and Smith , L. (eds.). (1993). *A Dynamical Systems Approach to the Development of Cognition and Action* . Cambridge, MA : MIT Press .

Thigpen , N. N. , and Keil , A. (2017). Event-related potentials. Reference Module in Neuroscience and Biobehavioral Psychology. Elsevier. [online resource]

Thomas , J. B. , Brier , M. R. , Bateman , R. J. , Snyder , A. Z. , Benzinger , T. L. , Xiong , C. , … Ances , B. M. (2014). Functional connectivity in autosomal dominant and late-onset Alzheimer disease . *JAMA Neurology* , ***71*** , 1111 – 22 .

Thompson , S. P. , and Newport , E. L. (2007). Statistical learning of syntax: The role of transitional probability . *Language Learning and Development* , ***3*** , 1 – 42 .

Tolman , E. C. (1948). Cognitive maps in rats and men . *Psychological Review* , ***55*** , 189 – 208 .

Tolman , E. C. , and Honzik , C. H. (1930). " Insight " in rats . *University of California Publications in Psychology* , ***4*** , 215 – 32 .

Tolman , E. C. , Ritchie , B. F. , and Kalish , D. (1946). Studies in spatial learning, II: Place learning versus response learning . *Journal of Experimental Psychology* , ***36*** , 221 – 9 .

Tononi , G. , and Koch , C. (2008). The neural correlates of consciousness . *Annals of the New York Academy of Sciences* , ***1124*** , 239 – 61 .

Trappenberg , T. (2010). *Fundamentals of Computational Neuroscience* (2nd ed.). Oxford : Oxford University Press .

Trauble , B. , Marinovic , V. , and Pauen , S. (2010). Early theory of mind competencies: Do infants understand others ′ beliefs? *Infancy* , ***15*** , 434 – 44 .

Trevethan , C. T. , Sahraie , A. , and Weiskrantz , L. (2007). Form discrimination in a case of blindsight . *Neuropsychologia* , ***45*** , 2092 – 103 .

Tsotsos , J. K. (2011). *A Computational Perspective on Visual Attention* . Cambridge, MA : MIT Press .

Tunney , R. J. , and Shanks , D. R. (2002). A re-examination of melioration and rational choice . *Journal of Behavioral Decision Making* , ***15*** , 291 – 311 .

Turing , A. M. (1936 – 1937). On computable numbers: With an application to the Entscheidungsproblem [Decision Problem] . *Proceedings of the London Mathematical Society* , ***42*** , 3 – 4 .

Turing , A. M. (1950). Computing machinery and intelligence . *Mind* , ***59*** , 433 – 60 .

Tye , M. (1991). *The Imagery Debate* . Cambridge, MA : MIT Press .

Tzafestas , S. G. (2018). Mobile robot control and navigation: A global overview . *Journal of Intelligent & Robotic Systems* , ***91*** , 35 – 58 .

Ullman , S. (2019). Using neuroscience to develop artificial intelligence . *Science* , ***363*** , 692 – 3 . doi: 10.1126/science.aau6595

Umilta , M. A. , Kohler , E. , Gallese , V. , et al. (2001). I know what you are doing: A neurophysiological study . *Neuron* , ***31*** , 155 – 65 .

Ungerleider , L. G. , and Mishkin , M. (1982). Two cortical visual systems . In D. J. Ingle , R. J. W. Mansfield , and M. A. Goodale (eds.), *Analysis of Visual Behavior* . Cambridge, MA : MIT Press .

Vaina , L. M. (ed.). (1991). *From the Retina to the Neocortex* . Boston : Springer .

Van den Bussche , E. , Hughes , G. , Humbeeck , N. V. , and Reynvoet , B. (2010). The relation between consciousness and attention: An empirical study using the priming paradigm . *Consciousness and Cognition* , ***19*** , 86 – 9 .

Van Essen , D. C. , and Gallant , J. L. (1994). Neural mechanisms of form and motion processing in the primate visual system . *Neuron* , *13* , 1 – 10. Reprinted in Bechtel, W., Mandik, P., Mundale, J., and Stufflebeam, R. S. (eds.). (2001). *Philosophy and the Neurosciences: A Reader* . Malden, MA : Blackwell .

Van Gelder , T. (1995). What might cognition be, if not computation? *Journal of Philosophy* , ***92*** , 345 – 81 .

Van Gelder , T. (1998). The dynamical hypothesis in cognitive science . *Behavioral and Brain Sciences* , ***21*** , 615 – 28 .

van Honk , J. , Terburg , D. , Thornton , H. , Stein , D. J. , & Morgan , B. (2016). Consequences of selective bilateral lesions to the basolateral amygdala in humans . In D. G. Amaral & R. Adolphs (eds.), *Living without an amygdala* . New York : Guilford Press .

Veselinovi ć , T. , Rajkumar , R. , Amort , L. , Junger , J. , Shah , N. J. , Fimm , B. , and Neuner , I. (2021). *Connectivity patterns in the core resting-state networks and their influence on cognition . Brain Connectivity* . Advance online publication . doi: 10.1089/brain.2020.0943

Voyer , D. , Voyer , S. , and Bryden , M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables . *Psychological Bulletin* , ***117*** , 250 – 70 .

Wagner , A. D. , Schacter , D. L. , Rotte , M. , Koutstaal , W. , Maril , A. , Dale , A. M. , … Buckner , R. L. (1998). Building memories: Remembering and forgetting of verbal experiences as predicted by brain activity . *Science* , ***281*** , 1188 – 91 .

Wang , S.-H. , and Baillargeon , R. (2008). Detecting impossible changes in infancy: A three-system account . *Trends in Cognitive Sciences* , ***12*** , 17 – 23 .

Warrington , E. , and Taylor , A. M. (1973). The contribution of the right parietal lobe to object recognition . *Cortex* , ***9*** , 152 – 64 .

Warrington , E. , and Taylor , A. M. (1978). Two categorical stages of object recognition . *Perception* , ***7*** , 695 – 705 .

Warwick , K. (2012). *Artificial Intelligence: The Basics* . London : Routledge .

Watson , J. B. (1913). Psychology as the behaviorist sees it . *Psychological Review* , ***20*** , 158 – 77 .

Waytz , A. , and Mitchell , J. (2011). Two mechanisms for simulating other minds: Dissociations between mirroring and self-projection . *Current Directions in Psychological Science* , ***20*** , 197 – 200 .

Webb , B. (1995). Using robots to model animals: A cricket test . *Robotics and Autonomous Systems* , ***16*** , 117 – 34 .

Weisberg , D. S. (2015). Pretend play . *Wiley Interdisciplinary Reviews: Cognitive Science* , ***6*** , 249 – 61 .

Weiskopf , D. A. (2004). The place of time in cognition . *British Journal for the Philosophy of Science* , ***55*** , 87 – 105 .

Wellman , H. M. (2018). Theory of mind: The state of the art . *European Journal of Developmental Psychology* , ***15*** , 728 – 55 . doi: 10.1080/17405629.2018.1435413

Wenzel , L. , Dörrenberg , S. , Proft , M. , Liszkowski , U. , and Rakoczy , H. (2020). Actions do not speak louder than words in an interactive false belief task . *Royal Society Open Science* , *7* , 191998. doi: 10.1098/rsos.191998

Westermann , G. , and Ruh , N. (2012). A neuroconstructivist model of past tense development and processing . *Psychological Review* , ***119*** , 649 – 67 .

White R. L. , III, and Snyder , L. H. (2007). Subthreshold microstimulation in frontal eye fields updates spatial memories . *Experimental Brain Research* , ***181*** , 477 – 92 .

Whyte , C. J. , and Smith , R. (2021). The predictive global neuronal workspace: A formal active inference model of visual consciousness . *Progress in Neurobiology* , ***199*** , 101918. doi: 10.1016/j.pneurobio.2020.101918

Wicker , B. , Keysers , C. , Plailly , J. , Royet , J. P. , Gallese , V. , and Rizzolatti , G. (2003). Both of us disgusted in my insula: The common neural basis of seeing and feeling disgust . *Neuron* , ***40*** , 655 – 64 .

Wilson , R. A. (2008). The drink you have when you ′ re not having a drink . *Mind and Language* , ***23*** , 273 – 83 .

Wimmer , H. , and Perner , J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children ′ s understanding of deception . *Cognition* , ***13*** , 103 – 28 .

Winfield , A. F. T. (2012). *Robotics: A Very Short Introduction* . Oxford : Oxford University Press .

Winograd , T. (1972). *Understanding Natural Language* . New York : Academic Press .

Winograd , T. (1973). A procedural model of language understanding . In R. C. Schank and A. M. Colby (eds.), *Computer Models of Thought and Language* . San Francisco : W. H. Freeman .

Womelsdorf , T. , Schoffelen , J. M. , Oostenveld , R. , Singer , W. , Desimone , R. , Engel , A. K. , & Fries , P. (2007). Modulation of neuronal interactions through neuronal synchronization . *Science* , ***316*** , 1609 – 12 .

Woodward , A. , and Needham , A. (2009). *Learning and the Infant Mind* . Oxford : Oxford University Press .

Wu , X. , Kumar , V. , Quinlan , J. R. , et al. (2008). Top 10 algorithms in data mining . *Knowledge and Information Systems* , ***14*** , 1 – 37 .

Xu , F. , and Tenenbaum , J. B. (2007). Word learning as Bayesian inference . *Psychological Review* , ***114*** , 245 – 72 .

Xue , G. (2018). The neural representations underlying human episodic memory . *Trends in Cognitive Sciences* , ***22*** , 544 – 61 .

Yates , L. , and Hobson , H. (2020). Continuing to look in the mirror: A review of neuroscientific evidence for the broken mirror hypothesis, EP-M model and STORM model of autism spectrum conditions . *Autism* , ***24*** , 1945 – 59 . doi: 10.1177/1362361320936945

Zacks , J. M. (2008). Neuroimaging studies of mental rotation: A meta-analysis and review . *Journal of Cognitive Neuroscience* , ***20*** , 1 – 19 .

Zeki , S. M. (1978). Functional specialization in the visual cortex of the rhesus monkey . *Nature* , ***274*** , 423 – 8 .

Zelazo , P. D. , Moscovitch , M. , and Thompson , E. (eds.). (2007). *The Cambridge Handbook of Consciousness* . Cambridge : Cambridge University Press .

Zelditch , M. L. , and Goswami , A. (2021). What does modularity mean? *Evolution & Development* , ***23*** , 377 – 403 .

Zhang , X. , Pan , W.-J. , and Keilholz , S. (2019). The relationship between local field potentials and the blood-oxygenation-level dependent MRI signal can be non-linear . *Frontiers in Neuroscience* , ***13*** , 1126 .

Zhao , B. , and Sala , S. D. (2018). Different representations and strategies in mental rotation . *Quarterly Journal of Experimental Psychology* , ***71*** , 1574 – 83 . doi: 10.1080/17470218.2017.1342670

Zylberberg , A. , Dehaene , S. , Roelfsema , P. R. , and Sigman , M. (2011). The human Turing machine: A neural framework for mental programs . *Trends in Cognitive Sciences* , ***15*** , 293 – 300 .

Index

2.5D sketch 44 – 47
3D sketch 44 – 47

Abrams, Richard 304
access consciousness (A-consciousness) 304 – 306 , 314
action potentials 178 – 179
actions
  complex behaviors 16 – 17
  situated cognition 281
  vision for 301 – 303 , 314
activation functions 98 , 101 , 106 – 109
activation levels 97 – 98
ACT-R architecture 164 – 168 , 170
Adolphs, Ralph 266 – 267 , 323 – 324 , 340
affective science 321 – 328 *see also* emotions
affective space 321 – 323
agent architectures 152 – 156
algorithmic-level analysis (cognitive systems) 42 , 44 – 45 , 227
algorithms
  dynamical systems 118
  formal analysis of language 20 – 22
  machine learning 231 , 234 – 236

rule-following 82
SHRDLU model 31
theory of computation 17 – 20
Allen (robot) 285 – 288
AlphaGo 246 – 248 , 250
amygdala
damage to 266 – 267 , 330 – 333
fear response 266 , 330
further reading 340
in humans 330 – 338
in rodents 328 – 330
anatomical connectivity 174 – 177 , 189 – 190 , 192 , 330
anatomy of the brain 53 – 59 , 171 *see also* brain mapping
Anderson, David 323 – 324
anger 267 , 318 – 321
A-not-B error 123 – 126
anterior cingulate cortex (ACC) 336
anxiety responses 326 , 329 – 330
appraisal theories of emotions 323 – 325 , 340
architectures
ACT-R architecture 164 – 168 , 170
for artificial agents 152 – 156
evolutionary basis 162 – 164
modularity 156
from subsumption architectures to behavior-based robotics 285 – 292
arousal 322
artificial intelligence (AI) 30 – 31
agent architectures 152 – 156

autonomous vehicles 345 – 346
deep learning 239
GOFAI approach 274 – 275 , 279 – 281 , 286 – 288 , 293 – 294
reinforcement learning 246 – 248
the Russian room argument 92
situated cognition and biorobotics 278 – 285
strong and weak artificial intelligence 92
from subsumption architectures to behavior-based robotics 285 – 292
testing cognitive science 36 – 37
see also machine learning ; robots
artificial neural networks *see* connectionist networks
attention
brain mapping 185 – 187
consciousness and 309 , 315
contention scheduling 309
from data to maps 188 – 190
further reading 193
information processing 22 , 24 – 26
attention effect 183
attraction, basins of 120 – 121
auditory localization 24 – 25
autism 254 – 258 , 271 – 272
autoencoders 242 – 243
autonomous vehicles 345 – 346

Baars, Bernard 308 – 309
babies *see* infancy
backpropagation
learning algorithm 61 , 107

McClelland – Jenkins network 226
Baillargeon, Renée 217 – 218 , 257 – 258
balance-beam problem 224 – 226
Baron-Cohen, Simon 254 – 260
basins of attraction 120 – 121
Bayesian decision theory 150
Bayesianism 130 – 131
binocular rivalry case study 138 – 141
conditional probability 132 – 134
degrees of belief and subjective probability 131 – 132
further reading 150
language learning 207 – 212 , 214
neuroeconomics 141 – 148
perception as a Bayesian problem 136 – 141
Bayes ′ s rule 132 , 134 – 136 , 208
Bayes, Thomas 131
behavior-based robotics 285 – 292 , 294
behaviorism
conditioning 12
reaction against 12 – 17
reinforcement learning 246 – 248
belief, degrees of 131 – 132
Bengio, Yoshua 239 – 241
Bernoulli, Nicholas 142
Berti, Anna 300 – 301
binary Boolean functions 99 – 102
binocular rivalry 138 – 141
biorobotics 281 – 285 , 293 – 294

Bisiach, Eduardo 298
blindsight 298 – 301
blocked design, fMRI 65 – 66
Block, Ned 304 – 306 , 314
BOLD signal
from data to maps 188 – 190
functional magnetic resonance imaging (fMRI) 65 – 69 , 71
mapping the brain 180 – 181 , 188 – 189
neural correlates 69 – 71
oxygenated and deoxygenated blood 65
resting state 343
underpinnings 71
Boolean functions 99 – 102
bottom-up analysis, two-visual-systems hypothesis 51 , 59
brain
anatomy 53 – 54 , 172 – 177
biological plausibility of neural networks 108 – 109 , 112
in cognitive science 50 – 51
cognitive systems as functional systems 51 – 52
extending computational modeling to the brain 59 – 61
formal properties 84 – 85
Human Connectome Project 342
levels of organization 5 – 6
lexical processing 61 – 65
memory for visual events 65 – 69
neural correlates of the BOLD signal 69 – 71
neural prosthetics 343 – 344
neuroeconomics 141 – 148

resting state 343
semantic properties 84 – 85
theory of mind 263 – 266
two-visual-systems hypothesis 54 – 59
vocabulary 55 – 56
brain – computer analogy 85
brain mapping
emotions 327 – 328
fMRI studies 65 – 69
from data to maps 188 – 190
further reading 192 – 193
locus-of-selection problem 181 – 185
networks for attention 185 – 187
PET studies 61 – 65
structure and function 171 – 178
studying cognitive functioning 178 – 181
Brewer, James B . 67 – 69
Broadbent, Donald 22 , 24 – 26 , 28 , 181 – 182
Broca ′ s area 264 – 265
Brooks, Rodney 280 , 285 – 292

calculus 115
Carruthers, Peter 170
causation by content 84 – 85
central nervous system (CNS) 4 – 6 , 122 , 317 – 318
central processing 152 – 153 , 156 – 158
cerebral cortex 54 , 174 – 177
cerebrum 53 – 54
Chalmers, David 305 – 306

Changeux, Jean-Pierre 309
chatterbots 30 – 31
cheater-detection module 158 – 162
Chilausky, Richard 236 – 237
child development
  Bayesian language learning 207 – 212
  dynamical systems 121 – 127
  language-learning models 201 – 207
  linguistic categories 211 – 212
  metarepresentation, theory of mind, and autism 254 – 258
  neural network models of children ′ s physical reasoning 221 – 226
  pretend play 252 – 254
  pronouns 209 – 211
  tense learning 204 – 207 *see also* infancy
Chomsky, Noam 20 – 22 , 27 , 201
chunking 22 – 24
Church – Turing thesis 19 , 27
circumplex model of emotion 322 – 323
Clark, Andy 293
cocktail party phenomenon 24 – 25 , 28
cognitive-deductive systems 32
cognitive dimension of emotions 323 – 325
cognitive maps 15 – 16
cognitive modules 156 – 157 , 160 *see also* modularity
cognitive science
  as challenge 6 – 8
  interdisciplinarity 1 – 3 , 344 – 345
  law and 344 – 345

looking ahead 341
" space " of 6 – 8 *see also* neuroscience
cognitive science, development of discipline 29
interdisciplinary model of vision 41
language and micro-worlds 30 – 36
mental images and representation 36 – 41
cognitive science, prehistory 11 – 12
connections and points of contact 25 – 26
information-processing models in psychology 22 – 25
linguistics and the formal analysis of language 20 – 22
reaction against behaviorism in psychology 12 – 17
theory of computation and the idea of an algorithm 17 – 20
cognitive systems
agent architectures 152 , 154 – 169
anatomy of the brain and the primary visual pathway 53 – 59
brain mapping 178 – 181
as functional systems 51 – 52
common sense 346 *see also* folk physics
competitive networks 108 – 109
complex behaviors 16 – 17
computational governor 118 , 127
computational-level analysis 41 – 42 , 227
computational modeling
children ' s physical reasoning 222
extending computational modeling to the brain 59 – 61
SHRDLU model 31
computational neuroscience 96
computation, theory of 17 – 20 , 27

computer – brain analogy 85
conditional probability 132 – 134 , 208
conditioned stimulus (CS) 329
conditioning
  association and reinforcement 13 – 15
  behaviorism 12
  fear 328 – 330
connectionist modeling 73 – 74 , 96 , 112 , 204 – 207
connectionist networks (artificial neural networks)
  balance-beam problem 224 – 226
  children ' s physical reasoning 221 – 226
  computational power 113
  contribution and limitations of 227
  further reading 214
  modeling cognitive processes 59 – 60
  object permanence 222 – 224 , 230
  pattern recognition 60 – 61
connectivity of the brain
  anatomical 174 – 177 , 189 – 190 , 192 , 330
  functional vs. effective 189 – 190
  Human Connectome Project 342
  neurons 96
  visual systems hypothesis 58 – 59
consciousness
  first- and third-person perspective 295 – 296
  further reading 314 – 315
  the hard problem and the knowledge argument 305 – 306
  information processing without conscious awareness 296 – 301

purpose of 301 – 304
research approaches 311 – 312
theories of 306 – 311
two types of 304 – 305
contention scheduling 309
continuity constraint 219
convergence *see* perceptron-convergence rule
ConvNets (convolutional neural networks) 243 – 246
cooperation, evolution of 160 – 161
co-opted mechanisms 266 – 268
co-opted systems 263
Cosmides, Leda 159 – 162 , 170
covert attention 185
Cox, James 159 – 160
cricket phonotaxis 282
cross-lesion disconnection experiments 57
culture, constants across 318 – 321

Damasio, Anthony 308
Darwin, Charles 340
Darwinian modules 158 – 164
Decision Field Theory 123
decision trees 232 – 234
declarative knowledge 165 – 167
deep learning
autonomous vehicles 345 – 346
machinery of 242 – 246
meaning of 239
reinforcement 231 – 248

representation learning 237 – 241
visual cortex and 239 – 241
default mode network (DMN) 343
degradation, graceful 52
degrees of belief 131 – 132
Dehaene, Stanislas 308 – 309
della Porta, Giambattista 138
dendrites 96 – 97
Dennett, Daniel 311
deontic conditionals 158 – 160
dichotic listening experiments 24 – 25
difference equations 115
differential equations 115
digital information processing 39
dimensionality of emotions 323 – 325
dimensionality reduction 242
disgust 266 – 267 , 318 – 321
dishabituation paradigm 216 – 221
dissociation 298 – 301
distributed neural networks 109 , 113
domain specificity 156 , 160 , 162
domain-specific knowledge 164
domain-specific modules 164
dopamine 267
dopaminergic neurons 248
dorsal route (primary visual cortex) 54 – 56 , 58 – 59 , 240 , 303
double dissociation 298
Draine, Sean 304

drawbridge experiments 217 – 218 , 230
dualism 12
dyadic relations 260
dynamical systems
  application to child development 121 – 127
  and cognitive science 114 – 115 , 117 – 121
  further reading 128 – 129
  meaning of 115 – 117

early-selection models 182
Ebbinghaus illusion 303
EEG *see* electroencephalography
effective connectivity 189 – 190
effector systems 153
Ekman, Paul 318 – 321 , 337 , 340
electrical stimulation studies 330
electroencephalography (EEG) 178 – 179 , 182 – 185 , 192 – 193
ELIZA (chatterbot) 30 – 31
emergent behaviors 288 – 290
emotion detector, the (TED) 259 , 267
emotions
  affective science 321 – 328
  basic emotion theories 318 – 321
  fear in humans 330 – 338
  fear in rodents 328 – 330
  further reading 340
  paired deficits 267
  physical symbol systems 317 – 318
  research context 316 – 317

empathizing system, the (TESS) 260 – 261 , 267
entropy (information theory) 234
event-related fMRI 66 – 69
event-related potentials (ERPs) 179 , 182 – 185
evolution
  cheater-detection module 158 – 162
  of cooperation 160 – 161
  Darwinian modules 158 – 164
  emotional responses 326
  facial expressions 318 – 321
  mental architecture 162 – 164
evolutionarily stable strategy 160 – 161
excitatory synapses 97
exoskeletons 344
expected monetary value 141
expected utility theory 141 – 148
expert systems 232 – 236 *see also* machine learning
extrastriate cortex 184
eye-direction detector (EDD) 259
eye movements
  gaze alignment 143
  saccadic 143 – 145 , 185 – 187
  smooth pursuit 143
eyewitness testimony 344 – 345

facial expressions 266 , 318 – 321
false-belief task 252 , 254 – 258 , 265 – 266 , 271
false-photograph task 265 – 266
Fang, Fang 303

fear
basic emotion theories 318 – 321
in humans 330 – 338
in rodents 328 – 330
neuroimaging 332 – 338
paired deficits 266 – 267
fear conditioning 328 – 330 , 335
feature engineering 231 , 237 – 238
feature learning 238 – 239
feedforward networks 106 – 109
Feinstein, Justin 330 – 333
first-order predicate calculus 274
first-person perspective (cognitive science) 295 – 296
fitness (evolutionary) 162 – 164
fMRI *see* functional magnetic resonance imaging
Fodor, Jerry
further reading 230
language of thought 82 – 89 , 94 , 199 – 201
limitations of artificial neural networks 227 – 229
modularity of mind 152 – 153 , 156 – 158 , 170
folk physics 218 , 220 – 221 , 346
forced-choice tests 298 – 299
forebrain 53 – 54
free riders 158 – 160
Frith, Uta 254 – 258
Frost, Douglas 298 – 299
functional connectivity 189 – 190
functional magnetic resonance imaging (fMRI)

BOLD signal 65 – 69 , 71
event-related 66 – 69
fear conditioning 335
further reading 74 , 192 – 193
mapping the brain 180 – 181 , 188 – 189
functional neuroimaging
brain ' s resting state 343
from data to maps 188 – 190
fear 332 – 338
networks for attention 185 – 187
using fMRI 65 – 69
using PET 61 – 65

Gage, Phineas, lesion study 327 – 328
Gahl, Susan 210
Game theory 161
gaze-alignment movements 143
General Problem Solver (GPS) program 78
genetics, use in affective science research 326 – 327
Gestalt principles 137 – 138
Glimcher, Paul 143 – 148
global neuronal workspace 309
global workspace theory of consciousness 308 , 315
goal-based agents 154 – 156
goal-completion mechanisms 317
Goel, Vinod 264
GOFAI (good old-fashioned artificial intelligence) 274 – 275 , 279 – 281 , 286 – 288 , 293 – 294
Goodale, Melvyn 301 – 303

Google Deep Mind 246 – 248 , 250
Gopnik, Alison 221
Gorman, Paul 60 – 61
graceful degradation 52
grammar
  formal analysis of language 20 – 22
  language-learning models 197 – 198
  tenses 202 – 207
Greenwald, Anthony 304
Griggs, Richard 159 – 160

Halting Problem 17 – 19
Hamilton, W. D . 163 – 164
happiness 318 – 321
Haugeland, John 279
Hebb, Donald 102 – 103
Hebbian learning 102 – 103 , 110 – 111
Held, Richard 298 – 299
hemiagnosia *see* unilateral spatial neglect
hemineglect *see* unilateral spatial neglect
Herrnstein, Richard 148
He, Shen 303
heuristic search 81
hidden layers 106 – 109
hidden units, modeling cognitive processes 60 – 61 , 105
hierarchical information processing 240 , 246
higher-order theories of consciousness 306 – 308 , 314 – 315
hindbrain 53 – 54
Hinton, Geoffrey 239 – 241

hippocampus 97 , 344
historical context *see* cognitive science , prehistory
Honzik, C. H . 13 – 15
Human Connectome Project 342
Human Genome Project 342
hybrid architectures 164 – 168
hypervigilance 332 – 333
hypothesis of subconscious information processing 17
hypovigilance 333

ID3 231 , 234 – 236
ImageNet competition 243 – 244
imagery debate 36 – 41 , 48 – 49
implementational-level analysis (cognitive systems) 42 , 227
infancy
  dishabituation 216 – 221
  folk physics 218 , 220 – 221
  joint visual attention 260
  mindreading 259 – 260
  motor control 121 – 123
  object permanence 123 – 126 , 216 – 224
  pretend play 252
inference by analogy 268 – 269
inferior frontal gyrus (IFG) 267 – 268
inferior parietal lobule (IPL) 267 – 268
inferior temporal cortex (ITC) 240
inferotemporal cortex 183 – 184
informational encapsulation 156
information, and representation 25 – 26

information gain (algorithms) 234
information processing
  complex behaviors 17
  consciousness 296 – 301
  dynamical systems theory 114 – 115
  imagery debate 39 – 41
  language of thought hypothesis 82 – 89
  Marr ′ s model of vision 41
  modularity 152 – 153 , 156 – 158 , 164 – 168
  PET studies 61 – 65
  selective attention 26
  storage and 109 – 110
  subconscious 40
  theory of mind 263 – 266
information-processing models
  neural networks 96 – 99 , 109 – 111
  in psychology 22 – 25
information storage 109 – 110
inhibitory synapses 97
innatism 199 , 201
inner sense 261 – 262 , 307
insects
  affective science 325
  biorobotics 281 – 285
insula 266 – 267 , 335
integration, principle of 178
intelligence
  physical symbol systems and problem solving 79 – 80

physical symbol systems and Turing machines 82
Russian room argument 91
situated cognition and knowledge representation 280
intentionality detector (ID) 259
intentional realism 84
interdisciplinarity
cognitive science 1 – 3 , 344 – 345
psychology and neuroscience 3 – 6
Sloan Foundation 1978 report 1 – 3
interdisciplinary model of vision 41
intermediate-level actions (ILAs) 276 – 277
interocular suppression 303
interrupt mechanisms 317 – 318
introspection 261 – 262 , 307
intuitive knowledge 345 – 346
invariance *see* selectivity/invariance problem
invariance under translation (ConvNets) 245 – 246
isotropic processing 157

Jackson, Frank 306 , 314
James, William 216
Jasper, Herbert 143 – 148
Jenkins, E . 225 – 226
joint visual attention 260

Kalish, D . 15 – 16
Kanwisher, Nancy 265 – 266
Kelly, William 269
Kerszberg, Michel 309

kin selection 163 – 164
knockout experiments on emotions 326
knowledge
  ACT-R architecture 164 – 168
  autonomous vehicles 345 – 346
  massive modularity 164
  neural networks 109 , 113
  object permanence 123 – 126
  probabilistic 137 – 138
  SHRDLU model 31
  situated cognition and knowledge representation 280
knowledge argument, consciousness 306 , 314
knowledge-how 166
knowledge-that 166
Kosslyn, Stephen 40 – 41
Kuczaj, Stan 202

language
  development of cognitive science 30 – 36
  formal analysis of 20 – 22
  grammatical rules 197 – 199
  humans compared to animals 30
  understanding and learning 198 – 199
language-learning models 197 – 198
  Bayesianism 207 – 212 , 214
  further reading 214
  grammatical rules 197 – 198
  language of thought 199 – 201
  linguistic categories 211 – 212

neural networks 201 – 207
pronouns 209 – 211
semantic priming 297
tenses 204 – 207 , 214
word segmentation 208 – 209
language of thought hypothesis 82 – 89 , 94 , 199 – 201
Lashley, Karl 16 – 17 , 275 – 278
latent learning 14 – 15
lateral geniculate nucleus (LGN) 240
late-selection models 182
law, cognitive science and 344 – 345
Lazarus, Richard 324
learning
neural networks 110 – 111
in single-layer networks 102 – 103
without reinforcement 13 – 15
learning agents 154 – 156
learning algorithms 108
LeCun, Yann 239 – 241
lesion studies 327 – 328 , 330
Leslie, Alan
autism studies 254 – 258
false-belief task 254 – 258
further reading 271
pretend play 252 – 254
lexical access 63
lexical processing 61 – 66 , 190
Lidz, Jeffrey 210

light intensity 45 – 46
limbic system 266
linear functions 98
linear separability 103 – 105
linguistic categories 211 – 212
linguistic processing 63
linguistics 20 – 22 , 27 *see also* language-learning models
listening 24 – 25
lobes 54 , 56
local algorithms 108
local field potential (LFP) 71
localist networks 109 , 113
locus-of-selection problem 181 – 185
Logothetis, Nikos 71
low-level actions (LLAs) 276 – 277
Luria, Alexander 172 – 174
Luzzatti, Claudio 298

machine learning 231 – 232
  deep-reinforcement learning 231 – 248
  expert systems 232 – 236
  further reading 250
  machinery of deep learning 242 – 246
  representation learning and deep learning 237 – 241
magnetoencephalography (MEG) 179
make-believe *see* pretend play
mapping functions 99 – 105
mapping the brain *see* brain mapping
Marchman, Virginia 206 – 207

Marcus, Gary 207
Marr, David 41 , 49 , 227
masked priming 304 , 314
massive modularity hypothesis 158 – 164
Matarić , Maja 289 – 292
mathematics
computation, theory of 17 – 20 , 27
information processing 22
McClelland, Jay 59 – 60 , 204 , 222 , 225 – 226
means ‐ end analysis 81
medial prefrontal cortex (MPFC) 269 , 337
MEG *see* magnetoencephalography
melioration theory 147 – 148
Meltzoff, Andrew 221
memory
dynamical systems 125 – 126
functional neuroimaging 61 – 65
modularity 165 – 167
visual events 61 – 69
visuospatial attention 186 – 187
memory span 24 – 25
mental images 36 – 41
mental rotation 37 – 38
metarepresentation 253
autism and theory of mind 254 – 258
pretend play and 252 – 254
Metzler, Jacqueline 37 – 38
Michalski, Ryszard 236 – 237

microelectrode recording 69 – 71
micro-worlds 31 – 36
midbrain 53 – 54
Miller, George 22 – 24
Milner, David 301 – 303
mindreading 251 – 252
  cognitive neuroscience 263 – 269 , 272
  false-belief task 254 – 258
  first steps in 259 – 260
  further reading 271 – 273
  link to pretend play 252 – 253
  as simulation 261 – 263
  as system 258 – 261
mirror neurons 178 – 180 , 267 – 268 , 272
mirror-neuron system (MNS) 267 – 268
Mishkin, Mortimer 51 , 53 – 59
Mobbs, Dean 335 – 337
modularity
  ACT-R architecture 164 – 168
  consciousness and 308
  Fodor on the modularity of mind 152 – 153 , 156 – 158
  further reading 170
  information processing 156 – 158 , 297
  massive modularity hypothesis 158 – 164
monkey studies
  cross-lesion disconnection experiments 57
  experiments on expected utility 143 – 148
  mirror neurons 178 – 180 , 267 – 268

saccadic eye movements 185 – 187
morphological computation 281 – 285
mortgage expert systems 232 – 234
motor control 121 – 123
multiagent programming 290 – 292
multilayer neural networks 106 – 109
multiunit activity (MUA) 71
Munakata, Yuko 222 – 223
MYCIN 232

naive physics 218 , 220 – 221
nativism 199 , 201
neonates 216 *see also* infancy
Nerd Herd 290 – 292
nervous system 5 – 6
neural activation patterns 222 – 223
neural correlates of consciousness (NCC) 315
neural correlates of the BOLD signal 69 – 71
neural implementation 46 – 47
neural networks 73 – 74
biological plausibility 108 – 109 , 112
computational modeling 59 – 60
ConvNets 243 – 246
fear 334 – 335
further reading 112 – 113
information-processing models 96 – 99 , 109 – 111
language-learning models 201 – 207
models of children ' s physical reasoning 221 – 226
the question of levels 226 – 229

single-layer networks and Boolean functions 99 – 105
training multilayer networks 106 – 109
neural pathways 5 – 6
neural prosthetics 343 – 344
neurites 96 – 97
neuroanatomy 53 – 59 , 171 *see also* brain mapping
neuroeconomics 141 – 148
neuroimaging
brain ′ s resting state 343
fear 332 – 338
mapping the brain 180 – 181 , 188 – 190
networks for attention 185 – 187
using fMRI 65 – 69
using PET 61 – 65
neurons
activity of 97
expected utility 143 – 148
neurophilosophy 2 , 343 – 344
neuroscience
branches of 4 – 7
contrast with psychology 3 – 6
from data to maps 188 – 190
mindreading 263 – 269 , 272
studying cognitive functioning 178 – 181 *see also* brain mapping
Newell, Allen 78 , 93 – 94
Newport, Elissa 209
nonconscious priming 296 – 297 , 333
masked 304 , 314

nonconscious thought
  blindsight and unilateral spatial neglect 298 – 301
  vision for action and perception 301 – 303
nonlinear activation function 98 – 99
nonlinear dynamical systems 120 – 121
nonmodular processing 156 – 158 , 297

object identification 183 – 184
object permanence 123 – 126 , 216 – 224
object substitution 252
Onishi, Kristine 257 – 258
optogenetics 327
orbitomedial prefrontal cortex (omPFC) 336
overregularization errors 202 – 203 , 205 – 206

paired deficits 266 – 267
parahippocampal cortex 68 – 69
parallel processing 59 – 60 , 63 – 64
parietal cortex 55 – 56
partial volume effects 189
pattern-matching modules 167
pattern recognition 60 – 61 , 346
Pavlov ′ s dogs 13
perception
  as a Bayesian problem 136 – 141
  locus-of-selection problem 181 – 185
  situated cognition 281
  vision for 301 – 303 , 314
perceptron-convergence rule 102 – 103 , 112 – 113

limits of 103 – 105
perceptual recognition 43 – 44
perceptual rivalry 138 – 141
peripheral attention 188
Perner, Joseph 271
PET *see* positron emission tomography
pharmacogenetics 327
phenomenal consciousness (P-consciousness) 304 – 306 , 314
philosophy
neurophilosophy 2
schools of 3
phrase structures (language) 20 – 22
physical symbol systems
children ′ s physical reasoning 222
emotions 317 – 318
as hypothesis 78 – 82
infant folk physics 220 – 221
information storage and information processing 109 – 110
language-learning models 201
language of thought 82 – 89
limitations of artificial neural networks 227 – 229
modularity 164 – 168
Russian room argument 88 – 92
Turing Test 91
Piaget, Jean 123 – 126 , 216 – 218
Pinker, Steven 170 , 203 , 205 – 206
place learning 15 – 16
PLANEX (SHAKEY) 277 – 278

Platt, Michael 143 – 148
Plunkett, Kim 206 – 207
Pöppel, Ernst 298 – 299
position-invariant object recognition 108 – 109
positron emission tomography (PET)
  further reading 73 – 74
  lexical processing 61 – 65
  mapping the brain 180 – 181 , 188 – 189
  as process 62 – 63
posterior cingulate cortex 336
posterior parietal cortex 57 – 58
posterior probability 208
posttraumatic stress disorder (PTSD) 335
poverty-of-the-stimulus arguments 201
predicate calculus 86 – 88
premotor cortex 267
prestriate cortex 57 – 59
pretend play
  link to mindreading 252 – 253
  metarepresentation and 252 – 254
  significance of 252
primal sketch 44 – 47
primary visual cortex
  anatomy 54
  dorsal route 54 – 56 , 58 – 59 , 240 , 303
  removal of 57 – 58
  two-visual-systems hypothesis 53 – 59
  ventral route 54 – 59 , 240 , 303

vision for action and perception 301 – 303
priming 296 – 297
Prince, Alan 203 , 205 – 206
principle of cohesion 218 – 219
principle of contact 219
principle of segregation 174 – 177
prior probability 208
prisoner ′ s dilemma 161
probability
Bayesianism 130 – 131
Bayes ′ s rule 134 – 136
binocular rivalry case study 138 – 141
conditional 132 – 134 , 208
expected utility theory 141 – 148
language learning 207 – 212
neurons coding for 147 – 148
subjective probability 131 – 132
problem solving
General Problem Solver (GPS) program 78
intelligence 79 – 80
traveling salesperson problem 79 – 81
procedural knowledge 165 – 167
pronominal anaphora 209 – 211
pronouns 209 – 211
propagation *see* backpropagation learning algorithm
propositional attitudes 83 – 84
propositional logic 79
psychological states 83 – 84

psychology
behaviorism and the reaction against 12 – 17
contrast with neuroscience 3 – 6
schools of 3
subfields 3 – 4 , 7
psychophysics 24 – 25
Pylyshyn, Zenon 227 – 230

Quinean property 157
Quinlan, Ross 234 – 236

radical simulationism 262 – 263
rat studies
fear 328 – 330
learning without reinforcement 13 – 15
recurrent networks 223 – 224
reflex agents 153 – 156
reflex systems 152
Regier, Terry 210
representation
and mental images 36 – 41
metarepresentation and pretend play 252 – 254
metarepresentation, theory of mind, and autism 254 – 258
situated cognition and knowledge representation 280
representational primitives 44
representation learning 231 , 238 – 239
response learning 15 – 16
reverse engineering 117
reward prediction error hypothesis 248

Ritchie, B. F . 15 – 16
Rizzolatti, Giacomo 178 , 267 , 300 – 301
robotic neuroprosthesis 343 – 344
robot reply, the Russian room argument 91 – 92
robots
Allen 285 – 288
ELIZA 30 – 31
further reading 293
GOFAI 274 – 275 , 279 – 281 , 286 – 288 , 293 – 294
multiagent programming 290 – 292
SHAKEY 275 – 278 , 286 – 288 , 293
situated cognition and biorobotics 278 – 285
from subsumption architectures to behavior-based robotics 285 – 292
TOTO 288 – 290
WANDA 283 – 284
Yokoi hand 283 – 285 *see also* artificial intelligence (AI)
Rock, Irving 137
rodent studies
fear 328 – 330
knockout experiments on emotions 326
learning without reinforcement 13 – 15
Rumelhart, David 59 – 60 , 204
Russell, James G . 322 – 323
Russian room argument 88 – 92

saccadic eye movements 143 – 145 , 185 – 187
sadness 318 – 321
Sahraie, Arah 299 – 300
Saxe, Rebecca 265 – 266 , 337

scientific psychology 3 – 4
scotoma 298
search-space 79 – 81
Searle, John 88 – 92 , 94
Sejnowski, Terrence 60 – 61
selective attention
  further reading 193
  information processing 22 , 24 – 26
  locus-of-selection problem 181 – 185
selectivity/invariance problem 241
self-driving cars 345 – 346
self-reflection 268 – 269
self-sacrifice 163 – 164
semantic priming 297 , 300 – 301 , 304
semantic properties (the brain) 84 – 85
semantics
  language of thought hypothesis 86 – 88
  Russian room argument 91 – 92
semantic systems, SHRDLU model 32
sensory systems 153 – 156
sentences, propositional logic 79
sentence structure 20 – 22
serial order 16 – 17
serial processing 63 – 64
SHAKEY (robot) 275 – 278 , 286 – 288 , 293
Shannon, Claude E . 22
shared attention mechanism (SAM) 260
shared weights (ConvNets) 245

Shepard, Roger 37 – 38
Shepherd, Gordon 5 – 7
SHRDLU model 31 – 36 , 48 , 279
Siegler, Bob 225
sigmoid function 98 – 99
Simon, Herbert 78 , 93 – 94 , 317 – 318
simulation theory 252 , 261 – 263
single-layer neural networks 99 – 105
single-neuron recordings 96
single-unit recordings 182 – 185
single-word processing 63
situated cognition
  biorobotics 281 – 285
  the challenge of building a situated agent 279 – 281
  further reading 293
  knowledge representation 280
  as theory 278
Skinner boxes 13 – 14
sleep, brain activity during 178 – 179
Sloan Foundation 1978 report 1 – 3
Smith, Linda 122 – 123 , 125
smooth-pursuit eye movements 143
solidity constraint 219
sparse connectivity (ConvNets) 245
spatially selective attention 185 – 187
spatial neglect 58 , 298 – 301
spatial resolution 6 – 7
spatial working memory 186 – 188

Spelke, Elizabeth 218 – 221
spike density function (SDF) 71
standard simulationism 261 – 262
state space 115 – 117 , 120 – 121
stereotactic maps 189
striate cortex 183 – 184
STRIPS planner (SHAKEY) 277 – 278
strong artificial intelligence 92
subconscious information processing 17 , 40 *see also* nonconscious thought
subjective probability 131 – 132
subsumption architectures 285 – 292
subsymbolic dimension of knowledge 167 – 168
sulpiride 267
SuperVision 244
surprise 318 – 321
symbols *see* physical symbol systems
synapses 97 , 178 – 179
synsets 243 – 244
syntactic structures of language 20 – 22
syntactic systems, SHRDLU model 32
syntax
  formal analysis of language 20 – 22
  language of thought hypothesis 86 – 88
  Russian room argument 91 – 92

task analysis, complex behaviors 17
TED *see* emotion detector, the
temporal cortex 56

temporal resolution, neuroscience 6 – 7
Tenenbaum, Joshua 211 – 212
tenses (language) 202 – 207 , 214
TESS *see* empathizing system, the
Thelen, Esther 122 – 123 , 125
theory of mind
  metarepresentation and autism 254 – 258
  modularity 164
theory-of-mind mechanism (TOMM) 260 – 261 , 263 – 266
third-person perspective (cognitive science) 295 – 296
Thompson, Susan 209
threshold functions 98 – 99
TIT FOR TAT strategy 160 – 161
Tolman, Edward C . 13 – 16
Tooby, John 159 – 162 , 170
top-down analysis, vision 42 – 44
TOTO (robot) 288 – 290
transformational structures (language) 20 – 22
transitional probabilities 208 – 209
traveling salesperson problem 79 – 81
Trevethan, Ceri 299 – 300
triadic relations 260
truth condition (TC) 200
truth rule 200
truth tables 100
Turing, Alan 17 – 20 , 27
Turing machines 17 – 20 , 78 – 79 , 82
Turing Test 91

two-visual-systems hypothesis 51 , 53 – 59 , 240

unconditioned stimulus (US) 329
Ungerleider, Leslie 51 , 54 – 59
unilateral spatial neglect 58 , 298 – 301
Urbach – Wiethe disease 266 – 267 , 330 – 338
utility 141 – 148

valence 322
Van Gelder, Tim
  computational governor 118
  dynamical systems 117 , 127
  further reading 128 – 129
  Watt governor 118 – 121
ventral premotor cortex (PMC) 267 – 268
ventral route (primary visual cortex) 54 – 59 , 240 , 303
vision
  Marr ' s model 41 , 49
  memory for visual events 65 – 69
  primary visual cortex 54
  two-visual-systems hypothesis 51 , 54 – 59 , 240
vision for action 301 – 303 , 314
vision for perception 301 – 303 , 314
visual cortex
  anatomy 54
  deep learning 239 – 241
  dorsal route 54 – 56 , 58 – 59 , 240 , 303
  removal of 57 – 58
  two-visual-systems hypothesis 53 – 59

ventral route 54 – 59 , 240 , 303
vision for action and perception 301 – 303
visual form agnosia 302
visuospatial attention 185 – 187
vitalism 311 – 312
von Helmholtz, Hermann 137
voxel 189 , 265 – 266

WANDA 283 – 284
Warrington, Elizabeth 43 – 44
Wason selection task 158 – 160
Watson, John B . 334
Watt governor 118 – 122 , 127
Watt, James 117 – 118
weak artificial intelligence 92
Webb, Barbara 282
Weiskrantz, Larry 299 – 300
Wickelfeatures 204 – 205
Wimmer, Heinz 271
Winograd, Terry 30 – 31 , 48 , 279
word segmentation 208 – 209

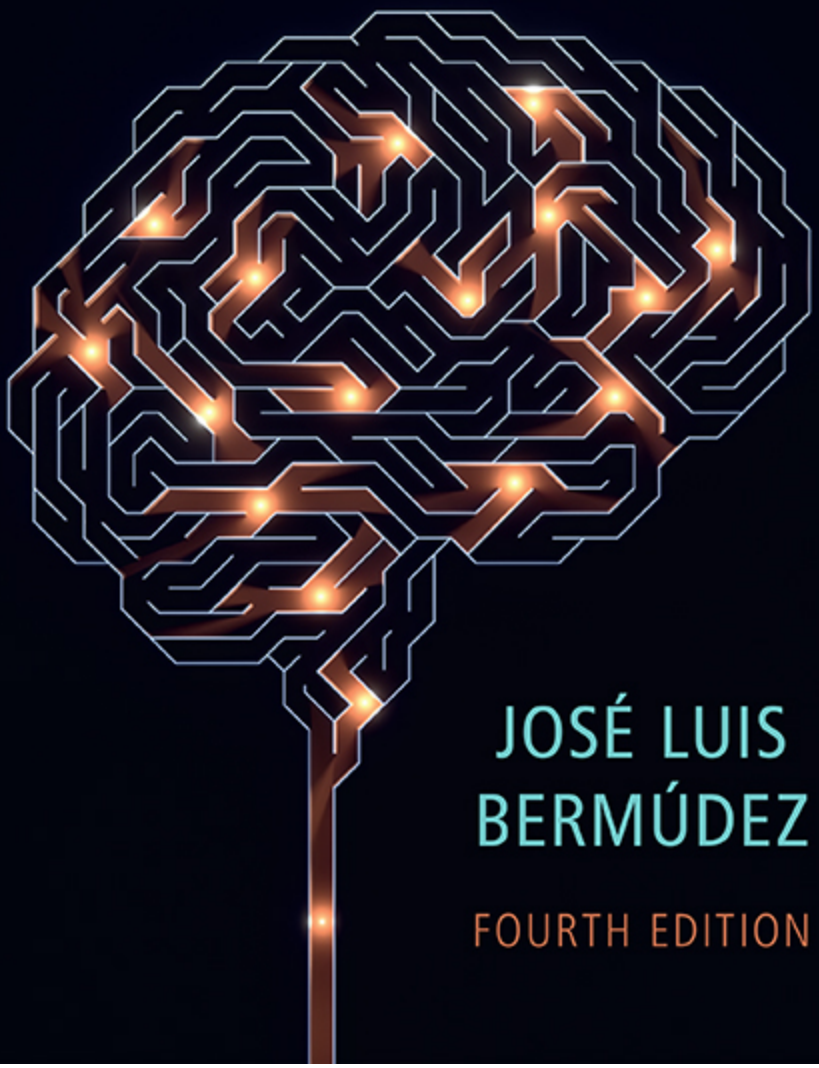
Xu, Fei 211 – 212

Yokoi hand 283 – 285

zero-crossings 46

An Introduction to the Science of the Mind

Cognitive Science



JOSÉ LUIS BERMÚDEZ

FOURTH EDITION