GOOD THINKING

Seven Powerful Ideas That Influence the Way We Think

BY DENISE D. CUMMINS

- Why would a psychologist label your idea a “creative insight”?
- What do economists mean when they refer to you as a “rational agent”?
- How can a philosopher still be logical while asking you to obey “moral imperatives”?

The answers are rooted in the critical–and sometimes counterintuitive–concepts that experts use to make decisions. Their methods determine whether we are guilty or innocent, where we should invest our money, and whether a drug effectively treats a particular illness.

Good Thinking: Seven Powerful Ideas That Influence the Way We Think explores the ways experts across various fields solve the problems that directly impact our lives. After reading this book, you will know how the best and brightest thinkers decide, argue, and tell right from wrong. But you will also see how our own “flawed” thinking can be used for the better.

DENISE D. CUMMINS is adjunct professor of psychology and philosophy at the University of Illinois at Urbana-Champaign and is the author or editor of three books, most recently *Minds, Brains, and Computers: The Foundations of Cognitive Science*. She lives in Champaign with her husband and two daughters.

Cover image: © haveseen / iStockphoto.com

Cover design by David Levy

CAMBRIDGE
UNIVERSITY PRESS
www.cambridge.org

ISBN 978-0-521-14550-3

CUMMINS GOOD THINKING

CAMBRIDGE

GOOD THINKING

Seven Powerful Ideas That Influence the Way We Think

DENISE D. CUMMINS

Good Thinking

After reading this book, you will know what the best and brightest thinkers have taught us about the best ways to decide, argue, solve, and judge right from wrong. You will also understand why analogy is the core of creative insight and genius, why we can so easily get taken advantage of in the stock market, why liberals and conservatives both feel confident that they are the more moral party, why you often cannot persuade people to change their beliefs in the face of a good argument, and why leaving an unsolved problem alone for a while makes it more solvable.

Denise D. Cummins is adjunct professor of psychology and philosophy at the University of Illinois at Urbana-Champaign. She joined the faculty in 2005. She has taught and conducted research at Yale University, the University of California, the University of Arizona, and the Max Planck Institute for Adaptive Behavior in Berlin. She is the co-editor of *Minds, Brains, and Computers: The Foundations of Cognitive Science*, and *The Evolution of Mind* and author of *The Other Side of Psychology: How Experimental Psychologists Find Out About the Way We Think and Act*. Her publications also include dozens of articles in scholarly journals such as the *Journal of Experimental Psychology*, *Cognition*, and *Synthese*. She has been an invited speaker at such prestigious institutions as St. Andrews College (Scotland), Durham University, Emory University, and Dartmouth College.

Good Thinking

Seven Powerful Ideas That Influence the Way We Think

Denise D. Cummins

University of Illinois

Cambridge
University Press

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Mexico City

Cambridge University Press
32 Avenue of the Americas, New York, NY 10013-2473, USA

www.cambridge.org
Information on this title: www.cambridge.org/9780521145503

First published 2012

Printed in the United States of America

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

Library of Congress Cataloging in Publication data
Cummins, Denise D.
Good thinking : seven powerful ideas that influence the way
we think / Denise Cummins.
p. cm
Includes bibliographical references and index.
ISBN 978-0-521-19204-0 – ISBN 978-0-521-14550-3 (pbk.)
1. Thought and thinking. 2. Game theory. 3. Rational choice theory. I. Title.
BF441.C86 2012
153.4–dc23 2011048651

ISBN 978-0-521-19204-0 Hardback
ISBN 978-0-521-14550-3 Paperback

Contents

List of Figures — *page* viii
List of Tables — ix
Acknowledgments — xi

ONE
Introduction — 1

TWO
Game Theory: When You Are Not the Only One Choosing — 5
The Basics of Game Theory — 6
Game Theory and the Battle of the Sexes — 8
Game Theory and Avoiding Mary's Inappropriate Co-Worker — 11
Game Theory and Deciding Whether or Not to Trust — 12
Experimental Economics: What Do People Actually Do? — 15
Differences in Power and Status Influence How Fairly We Treat Others — 19
Neuroscience Makes It Clearer Why We Behave the Way We Do — 23
The Evolution of Cooperation — 26

THREE
Rational Choice: Choosing What Is Most Likely to Give You What You Want — 31
How the "Big Boys" Think about Decision Making — 31
How to Be More Bayesian — 41
When We Are Not Bayesian — 43
How the Question Is Framed Determines Whether You Get It Right or Wrong — 47

Your Brain on Decision Making 53
Decision Making in the Real World: The Economic Meltdown of 2008 56

FOUR
Moral Decision Making: How We Tell Right from Wrong 59
Church and State Weigh in on Morality 60
What David Hume Had to Say 62
What Immanuel Kant Had to Say 65
What Jeremy Bentham and John Stuart Mill Had to Say 70
What Seems Right to Us: The Psychology of Moral Judgment 72
Yes, but What Is Morality for? 77

FIVE
The Game of Logic 80
A Journey into Logic Land 84
Just How Logical Are People, Really? 91
What to Do When the World (or Your Mind) Changes 94

SIX
What Causes What? 99
The Paradox of Causality 99
How the Experts Decide What Causes What 101
How Your Brain Decides What Causes What 105
What Is Necessary? What Is Sufficient? 106
How What You Believe Influences How You Decide 109
The Causal Paradox Revisited: What Infants Told Us 111

SEVEN
Hypothesis Testing: Truth and Evidence 115
Confirmation Bias: Tell Me I'm Right 116
A More Realistic Study of Confirmation Bias 117
When Your Brain Is Biased 119
Science: How We Got Here 121
Prove I'm Wrong, or Give Me the Most Bang for the Buck? 122
OK, Show Me I'm Wrong 127
Stopgaps and Backup Systems 134

EIGHT
Problem Solving: Turning What You Don't Want into What You Want 138
When Problems Are Well-Defined 140
When Problems Are Not So Well-Defined 145
Finding the Way There 150
Artificial Intelligence: Machines That Think 152
How Experts Solve Problems 155
Insight and Genius 158

NINE
Analogical Reasoning: This Is Like That 168
Analogy as It Should Be Done 170
Analogy: How It Is Actually Done 173
Why Analogy Is the Core of Cognition 178

References 185
Index 197

Figures

2.1. The Battle of the Sexes in extensive form *page* 10
2.2. Prisoner's Dilemma game in extensive form 14
6.1. Philosopher David Lewis's counterfactual treatment of causality 108
7.1. The "Bell" curve for normally distributed data 130
8.1. An English sentence parse tree for a complex sentence constructed from four simple rules 145
8.2. When doing a breadth-first search, all the possibilities at a given level are searched before going to a new level 147
8.3. When doing a depth-first search, all the possibilities connected to lower levels are searched 147
8.4. How one problem solver searched for a solution 148
8.5. How memory is automatically searched, sometimes without awareness 162

Tables

2.1. Battle of the Sexes Game in Normal Form *page* 9
2.2. Matching Pennies Game in Normal Form 11
2.3. Prisoner's Dilemma Game in Normal Form 13
3.1. Statistics for Breast Cancer and Mammograms According to Patients' Age 34
3.2. Prevalence of Breast Cancer According to Age of Patient 41
3.3. The Relationship between Performance Confidence and Performance Accuracy 51
7.1. The Logic of Hypothesis Testing 131

Acknowledgments

It's pretty easy to figure out who to thank for all of this knowledge in my head: the luminaries across the ages whose ideas forever changed the way we think and live. I hope I have done justice to those I present in this book. And if you have not encountered these great minds and their profound insights yet, I hope this book will change the way you think about yourself and the idea of thought itself.

Thanks also to my editors – Simina Calin, for having faith in this project, and Adina Berk, for seeing it through to the end.

Closer to home, many thanks to Rob, Katya, and Masha for grounding me in heart and home so that I don't spend all of my time up in my head.

Box 2.1 appearing on pages 16–17 is reprinted here with kind permission from Springer Science+Business Media, originally appearing in "Intuition in insight and noninsight problem solving," by Janet Metcalfe, *Memory & Cognition*, volume 15, issue 3, January 1987.

ONE

Introduction

After two decades of teaching bright and curious university students, I came to a disturbing conclusion: Despite our best efforts to expose students to the ideas and insights that profoundly shape the way we think and live, most students were still pretty insulated within their particular disciplines. The science majors knew all about hypothesis testing but didn't know the first thing about moral theory. The philosophy and pre-law majors knew all about argumentation but didn't know the first thing about scientific investigation. Outside of the business school, precious few students knew anything about decision theories that drive the equity market and underlie economic policies that impact their lives – right down to whether or not they can get student loans. And outside of the psychology majors, virtually none knew that the way the brain is wired shapes the way we think, act, and feel. And then these bright and well-educated people take jobs as policy-makers, writers, scientists, lawyers, and teachers – bumping about in life with holes where some crucial bits of knowledge ought to be.

Does this really matter? Well, consider a Colorado DUI case that ended in acquittal in spite of overwhelming evidence. "It made no sense," the prosecutor complained. "It was an open-and-shut case. The guy's blood alcohol level was over the limit, he couldn't walk a straight line, and there were open beer cans in the car with his fingerprints on them." So why was the defendant acquitted? "I talked to one of the jury members after the verdict," the attorney reported, "and he said there was an astrologer among them. She cast a chart and argued that, according to the chart, the defendant couldn't have been driving drunk that day. So they couldn't get a majority past reasonable doubt," he sighed.

Do astrological charts constitute reasonable doubt? For most of us, the answer would be "no." But explaining exactly *why* we believe this – and why this jury's decision is troubling – is a challenge. We simply know that most of us would never get on a plane, drive across a bridge, obey a law, or concern ourselves with presidential elections if they were based on this type of reasoning and evidence. Instead, we readily do these things because we assume that planes, bridges, laws, and our system of government are the outcomes of a painstaking process of reasoning, evidence evaluation, and learning from past mistakes. We believe that reason is the steel thread that makes the fabric of our jurisprudence fair, our science accurate, and our social institutions able to withstand change. In short, we hold a core belief that action should be governed by reason.

Coextensive is the core belief that decisions we make while in the throes of heated emotion are likely to be bad ones, and those we make in the cold and clear light of reason will be better. This just seems self-evident. This Wikipedia entry succinctly captures our folk wisdom: "Reason is a way of thinking characterized by logic, analysis, and synthesis. It is often contrasted with emotionalism, which is thinking driven by desire, passion or prejudice. Reason attempts to discover what is true or what is best."

So firmly entrenched are these beliefs that it often comes as a surprise to us to learn that not everyone thinks this way. For example, in *The Suicide of Reason: Radical Islam's Threat to the West* (2007), Lee Harris argues that

> [T]he West has cultivated an ethos of individualism, reason and tolerance, and an elaborate system in which every actor, from the individual to the nation-state, seeks to resolve conflict through words. The entire system is built on the idea of self-interest ... Our worship of reason is making us easy prey for a ruthless, unscrupulous and extremely aggressive predator and may be contributing to a slow cultural "suicide."

To thinkers like Harris, reason is what makes us weak, indecisive, and vulnerable. Reason is what ensnarls us in words and makes us slow to act. And there is sufficient evidence that human reasoning is frail and fallible. Scientists who study human reasoning and decision making have documented the alarming frequency with which we are prone to error.

The fallibility of human reasoning was not lost on our founding fathers, nor is it lost on scientists and policy-makers who still depend upon it to make decisions that impact millions of lives. As Ayaan Hirsi Ali puts it

> Enlightenment thinkers, preoccupied with both individual freedom and secular and limited government, argued that human reason is fallible. They understood that reason is more than just rational thought; it is also a process of trial and error, the ability to learn from past mistakes. The Enlightenment cannot be fully appreciated without a strong awareness of just how frail human reason is. That is why concepts like doubt and reflection are central to any form of decision-making based on reason. ("Blind Faiths," *New York Times* 1/6/08)

So how have these all-important modes of "doubt and reflection" been incorporated in our decision making? There are models of reason that dominate western thought. These are the "jewels in the crown" of our method of inquiry. Or, to borrow a term from philosopher Robert Cummins, they constitute our knowledge bridges – reasoning that takes us from what we already know to what we want to know.

The purpose of this book is to lay out each of these "knowledge bridges" in plain English so that educated readers can decide for themselves just how much or how little confidence we should have in our "worship of reason." These methods are

1. Rational choice: Choose what is most likely to give you what you want
2. Game theory: What to do when you're not the only one making choices
3. Moral judgment: How we tell the difference between right and wrong
4. Scientific reasoning, which consists of
   - Hypothesis testing: The search for truth by evaluating evidence
   - Causal reasoning: Explaining, predicting, and preventing events
5. Logic: The search for truth through argumentation

These are the main methods of inquiry and decision making that underlie the decisions we make in our everyday lives, in jurisprudence, in politics, in economics, and in science. Before we can decide whether reason can indeed be trusted, we need to understand the tools of the trade – how the "game of reasoned thought" – is played by the best.

In addition, there are two other modes of reasoning that merit discussion. Although not as well formalized as the previous four, they are ubiquitous in human and non-human cognition.

6. Problem solving: The search for solutions to unwanted situations
7. Analogical reasoning: The heart and soul of insight, discovery, and genius

One last thought must be kept in mind: However rational and flawless these methods may seem, they are not implemented on infallible hardware. Instead, these models are implemented by flesh-and-blood human reasoners, or more specifically, by their neural circuitry. To fully appreciate the whole package of reason, we must be conversant with the way such circuits operate in different circumstances to yield decisions. For this reason, this book will detail important findings from the new fields of decision neuroscience that are pertinent to each of these models of thought.

After reading this book, readers should be empowered to decide for themselves whether human reasoning is as frail or as strong, as dangerous or as benign, or as superfluous or as crucial as it has been made out to be.

TWO

Game Theory

WHEN YOU ARE NOT THE ONLY ONE CHOOSING

John and Mary are trying to decide how to spend their Friday evening. John prefers to stay in and play videogames. Mary prefers to go to a movie. But they both prefer to be together rather than apart. You can see the problem. Any way they choose, one or both will be unhappy. If they play videogames, John will be happy, but Mary will be bored. If they go to a movie, Mary will be happy, but John will be settling for his second choice. If they go their separate ways, both will be unhappy.

This is a much harder decision to make than it seems at first blush because each decision maker is not the only one choosing, and the outcome for each depends on what the other does, and they both know that. Let's follow Mary and John a bit more.

It's now Monday afternoon, and Mary is trying to avoid an annoying co-worker who keeps asking her out on a date even though he knows she's married. There are only two places to eat near her workplace, Subway Sandwich Shop and Starbucks. If she goes to Subway, and the co-worker goes there as well, she won't be able to avoid him. She will be miserable, but he will be delighted. The same thing will happen if they both end up at Starbucks. But if she goes to Subway, and he goes to Starbucks, she will be relieved, and he will be frustrated. Same thing if she goes to Starbucks, and he goes to Subway. So, once again, the outcome for each party depends on what the other person does, and they both know that.

Meanwhile, John is facing a dilemma of his own at work. He and a co-worker jointly botched a report in a major way, and it ended up costing the company they work for $100,000. Their boss is in a rage and plans to make the person responsible repay the company out of his own pocket. He meets with each man separately and demands to know who botched the report. If they blame each other, he will fine each of them $50,000. If only one blames the other, the person blamed will be fined $100,000, and the other will get off scot-free. If they both refuse to blame the other, then the boss will fine them each $25,000 and write off the remaining $50,000. John has to decide whether to blame his co-worker or to keep mum. His co-worker is facing the same dilemma, and they both know it. So this is a matter of trust, and what happens to both depends on what the other does.

These are the kind of choices we face frequently in life. To a mathematician, these kinds of problems are called *games*, and the optimal choices associated with them can be determined by *game theory*.

The Basics of Game Theory

Oskar Morgenstern and John von Neumann formulated the basic concepts behind game theory in their 1944 book *Theory of Games and Economic Behavior*. First, certain assumptions have to be made that we've already encountered when we learned about Bayesian decision making: Agents have preferences that can be ordered in terms of utility (satisfaction), and they act logically according to those preferences.

A game is a decision-making situation involving more than one player. Each player is trying to maximize his or her payoffs, but each player's actual payoff depends on what the other players do. Games are defined in terms of the set of participants playing, the possible courses of action available to each agent, and the set of all possible payoffs. In *constant-sum games*, the total payoff (sum of what everyone can get) is the same for all possible outcomes. Think of TV networks competing for viewership. If there are ten million viewers, and three million of them are watching NBC, that means the other networks are down three million viewers. If two million of them switch to ABC, ABC gains two million viewers, and NBC loses two million viewers. One player's gain is another's loss, and the sum of the payoffs is the same regardless of who wins viewership and who loses viewership. In a *zero-sum game* (a special type of constant-sum game), payoffs sum to zero. If I win $1, you lose $1. So the payoffs are plus one for me and minus one for you, and the sum of the payoffs is zero. In a *non-zero-sum game*, the sum of all payoffs could be negative or positive: Everyone could suffer, or everyone could benefit, but the sum of the suffering or benefit across all players is the same for all possible outcomes. For example, it could be that no matter how this game is played, the sum of all payoffs will be $50, and everyone will win something. That means that if it's just you and me, and I win $30, then you will win $20. Or it could be that no matter how this game is played, the sum of all payoffs will be minus $50, meaning that if it's just you and me, and I lose $30, then you will lose $20.

Games can be cooperative or non-cooperative. In *cooperative games,* players can form coalitions or alliances in order to maximize expected utility. Think of the difference between singles and doubles in tennis. Singles tennis is a non-cooperative game – the players play as individuals and vie to win the match. In doubles, the players play as teams each consisting of two players. The players on one team cooperate to beat the other team to win the match. Basketball, football, and soccer are all examples of cooperative games (which a friend of mine calls "coalitional ball-moving games"). Singles tennis, chess tournaments, and most videogames are *non-cooperative games*; a single individual vies to win the game against a human or computer opponent.

At each stage of the game, the players do something – they choose an action. There can be many outcomes to the game depending on the actions the players take. We can think of these actions as strategic. In a basketball game, players can play offensively or defensively. They can choose to execute a series of passes aimed at positioning the ball strategically. Some strategies lead to better outcomes for a given player than other actions. A player's *best response* is any strategy that yields the highest possible payoff. If you are a player or a coach, your best response is the strategy that is most likely to allow you to win the game.

When the game has reached a state of play in which no player can unilaterally improve the outcome of the game, the game is at *equilibrium*. Each player has adopted a strategy that cannot improve his outcome given the other players' strategies. For example, when one person or one team wins a tennis match, we say that the game has reached equilibrium. The winners can't do any better because they have won the match. The losers can't do any better because there are no more points to win or no more games to play in the match. There could also be a draw, as in a chess stalemate, when neither party can make a move that will improve his or her position. The game is over, but neither wins.

Contrast this with the situation described in the movie *A Beautiful Mind*: A group of guys enter a bar. They all see the sexiest woman in the bar, and they all want to go home with her. If they all compete for her, only one can win, all the other women will be offended and leave, and the rest of the men will go home lonely. But if the men switch strategies from pursuing the sexiest woman to pursuing other women, they increase their chances that they will all go home happy. In other words, the men can do better by switching strategies, and everyone knows that.

In 1950, John Nash formalized this idea for cooperative games. In *Nash equilibrium*, each player plays a best response and correctly anticipates that her partner will do the same. If each player has chosen a strategy, and no player can benefit by changing his or her strategy while the other players keep theirs unchanged, then the current set of strategy choices and the corresponding payoffs constitute a *pure-strategy Nash equilibrium*. To check whether there is a pure-strategy Nash equilibrium, all you have to do is check whether either player can do better by switching strategies.

Game Theory and the Battle of the Sexes

Let's return to the dilemmas faced by John and Mary. In the first one, John preferred videogames to movies, Mary preferred the opposite, and both preferred to be together rather than apart. This game is called the Battle of the Sexes, and it has a very interesting property: It has two pure-strategy Nash equilibria.

As we've described it, the Battle of the Sexes is a *simultaneous game* – that is, the players choose at the same time without knowing

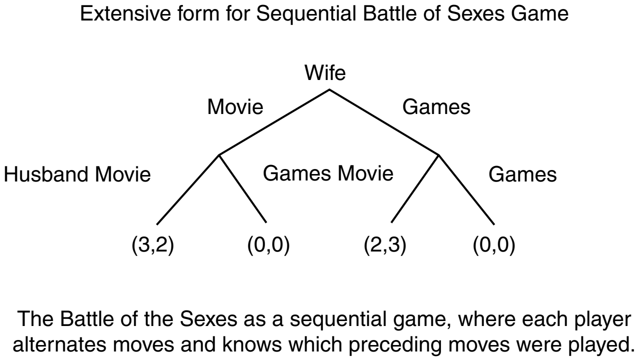
Table 2.1. *Battle of the Sexes Game in Normal Form*

	Mary	
John	Movies	Videogames
Movies	(3,2)	(0,0)
Videogames	(0,0)	(2,3)

what the others have chosen. Simultaneous games are represented using matrices that describe each player's move and payoff (and information). This is called *normal form*, and it constitutes a description of the strategies available to each player along with their payoffs. Table 2.1 presents the Battle of the Sexes game John and Mary face, represented in normal form.

Mary's best choice is movies, and John's is videogames – and they both know this. What if they both adopt their best choices? If Mary adopts her best choice (movies), then John knows he should switch strategies and choose to go to the movies as well. If John adopts his best choice (videogames), then Mary knows she should switch strategies and choose to stay home and play videogames. So there are two pure-strategy Nash equilibria here: movie-movie, and videogames-videogames. How do you break this deadlock?

One way to do this is for John and Mary to take turns – that is, let John have his top choice this time, and then let Mary have her top choice next time, and so on. The Battle of the Sexes then becomes a sequential game. In a sequential game, players alternate moves, knowing what choices have already been made. Suppose John and Mary write down whether they went to a movie or played videogames each time so that they both know where they stand in the game. If every player observes the moves of every other player who has gone before her, the game is one of *perfect information*. Suppose instead that they don't write it down, and Mary's memory is much better for this sort of thing than John's. If some (but not all) players have information about prior moves, the game is one of *imperfect information*. Sequential games are represented using *game trees* showing each move and each possible response along with payoffs (and information). This kind



The Battle of the Sexes as a sequential game, where each player alternates moves and knows which preceding moves were played.

FIGURE 2.1. The Battle of the Sexes in extensive form.

of representation is called *extensive form,* and it includes a complete description of the game, including the order of possible moves, payoffs, and information available to each player at each move. Figure 2.1 presents the Battle of the Sexes for John and Mary in extensive form.

What if John and Mary decide instead to break the deadlock by flipping a coin? If you introduce an element of chance into the game, it is called a *mixed-strategy equilibrium* game rather than a pure-strategy equilibrium game, and the best choice reduces to the probabilities associated with the element of chance introduced. Since Mary and John decided to flip a coin, for any given game, they both have a 50% chance that their preferred option will be chosen. Or they could play Rock, Paper, Scissors, adopting the preferred choice of the person who wins 2 out of 3 rounds. On each round, the chance of winning is 1 out of 3. If they decide to draw straws instead, with whoever draws the shortest straw winning, then the probability of getting one's choice is 1 out of the total number of straws.

Now here is the flash of brilliance of a beautiful mind: Nash proved that if there are a finite number of players and a finite number of strategies in a game, then there has to exist at least one Nash equilibrium, either pure strategy (choose a strategy and stick to it) or mixed strategy (introduce an element of chance). In 1994, the Nobel Prize in Economics was awarded to Nash, John Harsanyi, and Reinhard Selten for their work in game theory.

Table 2.2. *Matching Pennies Game in Normal Form*

Player 2	Player 1	
	Heads	Tails
Heads	(+1, −1)	(−1, +1)
Tails	(−1, +1)	(+1, −1)

Game Theory and Avoiding Mary's Inappropriate Co-Worker

Let's turn to the second problem described in the introduction where Mary was trying to avoid her inappropriate co-worker. This game is called *Matching Pennies*, and it also has a very interesting property: There are no pure-strategy Nash equilibria.

This game is called Matching Pennies because it has the same structure as the game you used to play as kids where you each have a penny in one hand and choose to put it on a table heads up or tails up. One of you wins if the coins match (two heads or two tails), and the other wins if they don't (one head and one tail). Let's say you win if they match, and your friend wins if they don't. If both of you play heads, then you will always win, and your friend will always lose. So your friend has an incentive to switch to playing tails. But that means that now you will always lose, so you have an incentive to switch to playing tails. That means your friend will now always lose, so your friend has an incentive to switch, and around and around you go. There is no pure-strategy Nash equilibrium for this game. Table 2.2 presents the game in normal form.

So what do you do? Probably what every kid who has ever played this game does: randomly switch between playing heads and tails. This gives each of you a 50% chance of winning.

As Nash proved, when you introduce an element of chance in a game that has no pure-strategy equilibrium, there is a mixed-strategy equilibrium. Now both players have a 50% chance of winning, so they might as well stick with what they're doing. So Mary might as well flip a coin to choose whether to go to Subway or Starbucks.

Game Theory and Deciding Whether or Not to Trust

Let's turn to the third situation: Should John cooperate with his co-worker and keep mum, or should he defect and rat him out? This game is called Prisoner's Dilemma because it has the kind of structure used to settle a case when there is not enough evidence to convict either party in a jointly committed crime. If you can get one of them to give evidence against the other, you can convict. The district attorney will try to make a deal, such as immunity, in exchange for evidence against the other guy. This game also has a very interesting property: *dominant-strategy equilibrium.*

In Prisoner's Dilemma, each player has a dominant strategy – that is, each party's best response does not depend on the strategies of the other players. No matter what the other player does, there is one plan that works best for each. If both rivals have dominant strategies that coincide, then the equilibrium is called a *dominant-strategy equilibrium*, a special case of a Nash equilibrium. In one-shot Prisoner's Dilemma, there are only two possible responses (cooperate or defect), and hence only two strategies (cooperate or defect). We can represent John and his co-worker's dilemma in normal form, as shown in Table 2.3.

Look at the top line. If John and his co-worker both rat each other out, they each have to pay $50,000. If John blames his co-worker, but his co-worker keeps mum, John pays nothing, and the co-worker pays the whole $100,000 himself. Now look at the bottom line. If instead John keeps mum, but his co-worker rats him out, then John will pay the whole $100,000 himself. If they both keep mum, then each will be responsible for only $25,000.

Both John and his co-worker are aware of the situation, and they have no control over what the other does. They each assume the other will do what gives him the best payoff. So the dominant strategy for each player in one-shot Prisoner's Dilemma is the same: *defect.*

Perhaps you think if they made a pact beforehand to keep mum, then the best choice (or the fairest choice) for John would be to just keep mum. But when it comes down to the wire, John has no idea whether his co-worker will honor that agreement or give into temptation. His best choice still remains the one that allows him to protect

Table 2.3. *Prisoner's Dilemma Game in Normal Form*

John	Co-worker	
	Defect	Cooperate
Defect	P = (–$50K, –$50K) Punishment for mutual defection	T= (0, –$100K) Temptation to defect
Cooperate	S = (–$100K, 0) Sucker's payoff for cooperating with a defector	R = (–$25K, –$25K) Reward for mutual cooperation

himself. The key point of game theory is that everyone knows the best strategy for each of the players, and everyone assumes the other players will play their best strategy.

What if John and his co-worker were good friends with a prior history of cooperation? Or if John and his co-worker are rivals with a prior history of intense competition? Does this change the nature of the game? In particular, does it change the strategy John should adopt – cooperate rather than defect – or does it not make any difference?

As it turns out, it matters a lot. If John has a history of Prisoner's Dilemma experiences with his co-worker, how he should behave depends on how his co-worker has behaved in the past. When Prisoner's Dilemma is played repeatedly, and everyone can keep track of previous decisions, it can be represented in extensive form shown in Figure 2.2.

In 1980, Robert Axelrod, a professor of political science at the University of Michigan, held a Prisoner's Dilemma tournament in which computer programs played games against one another and themselves repeatedly. The program that yielded the best outcome would be the winner of the tournament. Each program defined a strategy that specified whether to cooperate or defect based on the previous moves of both the strategy and the opponent. Some of the strategies entered in the tournament were:

- *Always defect*: This strategy defects on every turn. This is what game theory advocates. It is the safest strategy since it cannot be taken advantage of. However, it misses the chance to gain larger payoffs by cooperating with an opponent who is ready to cooperate.

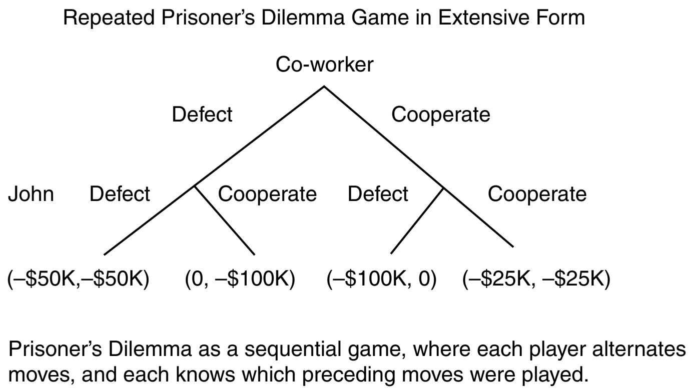

Repeated Prisoner's Dilemma Game in Extensive Form
Co-worker
Defect
Cooperate
John
Defect
Cooperate
Defect
Cooperate
(–$50K,–$50K)
(0, –$100K)
(–$100K, 0)
(–$25K, –$25K)
Prisoner's Dilemma as a sequential game, where each player alternates moves, and each knows which preceding moves were played.

FIGURE 2.2. Prisoner's Dilemma game in extensive form.

- *Always cooperate*: This strategy does very well when matched against itself. However, if the opponent chooses to defect, then this strategy will do poorly.
- *Random*: The strategy cooperates 50% of the time.
- *Tit-for-tat*: This strategy cooperates on the first move, and then does whatever its opponent has done on the previous move.

The first three strategies were prescribed in advance so they could not take advantage of knowing the opponent's previous moves and figuring out its strategy. That is, they didn't learn anything about their opponents as the tournament progressed. But the fourth, tit-for-tat, modified its behavior by looking back at the previous game – and only the previous game.

The results of the tournament were published in 1981 (Alexrod & Hamilton, 1981), and the winner was tit-for-tat. This strategy captures the full benefits of cooperation when matched against a friendly opponent, but does not risk being taken advantage of when matched against an opponent who defects. Note that tit-for-tat is a smart heuristic like those studied by Gigerenzer and his colleagues; it ignores part of the available information, involves practically no computation, and yet is highly successful. (See Chapter 3 for a discussion of Gigerenzer's smart heuristics.)

There were some other interesting results. When matched against itself, the tit-for-tat strategy always cooperates. That is because it always cooperates in the first game, so the outcome of the first game is cooperate-cooperate. When paired against a chronic defector, tit-for-tat cooperates on the first move and then defects forever after. When paired with a mindless strategy like random, tit-for-tat sinks to its opponent's level. For that reason, tit-for-tat can't be called a "best" strategy, but it was powerful enough to win the tournament.

So what should John do? If his prior history indicates that his co-worker is a chronic cooperator, he should cooperate. If instead, his co-worker is a chronic defector, he should defect. If his co-worker is unpredictable, he should defect. So the take-home message here is a player should try to figure out (or guess) his opponent's strategy and then pick a strategy that is best suited for the situation.

Experimental Economics: What Do People Actually Do?

So what happens when you make people play Prisoner's Dilemma? In experimental-economics studies, people play against other humans, and real money changes hands. When they leave the experiment, they go home with the payoffs they accrued while playing the game.

Remember that, according to standard game-theoretic analysis, a rational agent always (a) acts according to self-interest, (b) defines self-interest in terms of maximum payoff to oneself, (c) plays dominant strategy in games that have one, and (d) seeks Nash equilibrium in games that have one.

As we just saw, the rational choice in one-shot Prisoner's Dilemma is to defect. If you do this, an economist would consider you rational because you played the strategy that is most likely to bring you maximum gains. So you may be surprised to find that people cooperate in one-shot Prisoner's Dilemma games about 50% of the time (Camerer, 2003). When asked why, they don't list altruism as a reason. Instead, they typically say that they expect others to cooperate as well. They might say something like, "She'd do the same for me." This is consistent with the principle of reciprocity ("I'll help you if you'll help me in return") rather than unilateral altruism ("Go ahead, you take all the money, I don't mind"). We can say, then, that people approach these games with a bias toward cooperation, an approach that economists find irrational (Kelley & Stahelski, 1970). In fact, people seem to approach these kinds of transactions with an implicit *norm* of cooperation.

Social norms are standards of behavior that are based on widely shared beliefs about how individual group members ought to behave in a given situation. They are usually expressed as conditional rules that lay out how a person is *permitted*, *obligated*, or *forbidden* to behave in a certain way in a specific situation. Usually, demand for a social norm arises when actions cause positive or negative side effects for other people. People comply with social norms voluntarily when the norm matches their goals (self-interest). They comply with social norms involuntarily when group interest conflicts with their self-interest, but others have coercive authority to enforce the norms or the opportunity to punish non-cooperators.

Consider how people behave in Public Goods experiments. Contributions are made to a common good (benefit) that all members can consume, even those who do not contribute anything. All well and good, except for the problem of free riding – taking benefits while contributing nothing yourself. In Public Goods situations, each member has an incentive to free ride on the contributions of others. This can be modeled as Prisoner's Dilemma where cooperation is conditional (cooperate if everyone else does), and defection is free riding. Just as

Box 2.1. How Public-Good Experiments Work

Game Components

- Groups consist of n individuals, where n is greater than 2.
- Each individual is given a monetary endowment E.
- Individuals decide how much of E they keep for themselves and how much they spend on a group project.
- The experimenter multiplies the total amount spent on the group project by a number, b, that is greater than 1 but smaller than n.

- The multiplied sum of the member's contribution constitutes the proceeds from the group project.
- These proceeds are then distributed equally among the n members.

Outcomes

If all members keep their endowments, they each earn E. If all contribute their endowments, the sum of contributions is nE,

- yielding an income of (b/n)nE = bE,
- which is greater than E for each group member.

Example

E = 20, b = 2, n = 4
Income: (b/n)nE = bE

If nobody contributes, each earns 20. If everybody contributes everything, each earns (2/4)4(20) = 2(20) = 40. If everybody contributes 5, each earns (2/4)4(5) = 10, plus 15 they kept for themselves = 25.

Prisoners' Dilemma is a special case of the Public Goods game with n = 2 and two available actions: contributing nothing (defect) or contributing everything (cooperate). Both players in Prisoner's Dilemma are better off if they defect (because b/n < 1) regardless of what the opponent does.

in experiments with Prisoner's Dilemma, the incidence of defection depends on whether or not defectors (free riders) are punished.

One such study was reported by Fehr and Gächter (2000). People were given tokens that could be exchanged for real money. In each trial of the experiment, they were given the opportunity to keep their money or contribute some or all of it to a group project. For each token they kept, they earned another token. For each token contributed to the project, everyone – even those who did not contribute – earned one-quarter of a token. At the end of the experiment, these tokens were converted into real money according to a publicly known exchange rate. A purely self-interested person would never contribute anything in this experiment. Contributions started at about 50% on the first trial. But free riding began to occur – people began earning tokens without contributing anything to the project. As play progressed, people contributed less and less, and, by the tenth trial, contributions had dried up entirely.

But then at the eleventh trial, people were given the opportunity to punish the free riders by imposing a penalty that had to be paid in tokens. Just as in two-person Prisoner's Dilemma experiments, the results were dramatic: Cooperation jumped back up to 60%, then steadily climbed to 100% by the twentieth trial. These results showed that people contribute less if free riding is tolerated, and, over time, contributions can cease entirely.

Just as in two-person Prisoner's Dilemma games, people in Public Goods experiments show a strong willingness to punish others in order to foster cooperation – even if it costs themselves something to do so. In a study by Fehr and Fischbacher (2004a), every time a free rider was punished, both the punisher and the punishee had to pay one token. It turned out that people were willing to pay up to two tokens to punish free riders. This really makes economists scratch their heads because it means the desire to punish norm violations is strong enough to overcome self-interest.

The importance of the threat of punishment also has a great impact on people's behavior in Trust Games. This type of game allows people to act like investment bankers, particularly those managing trust accounts. In one-shot Trust Games, the investor is given money and is invited to invest it with the trustee. They can transfer as much or as little as they want to the trustee, but the experimenter triples the amount transferred. This tripling simulates investment earning. So if you choose to give $1 to the trustee, the trustee actually gets $3. The trustee is free to return as much as he or she chooses, including nothing at all.

If trustees acted according to the principle of pure self-interest, they would keep all of the money. But they don't. Typically, investors transfer about half of their money to the trustee, and trustees return a little less than that to the investors. But what if investors were allowed to impose a penalty if the rate of return was less than a specified rate? Fehr and Rockenbach (2003) tested these conditions and found that trustees returned 50% more money if the rate allowed them to earn some money themselves, but they returned 67% less money if the rate cost them money. This is sometimes referred to as the *Norm of Just Punishment* (as in, "OK, I guess I deserved a little punishment for my greed, but that much punishment is unfair!") Fehr and Rockenbach interpreted their results to mean that when punishment is harsh or unfair, altruism declines.

So what we've learned so far is this: People's behavior in repeated games puzzles economists greatly because we reward cooperation more generously and punish defection more severely than is predicted by standard game-theoretic analyses (Weg & Smith, 1993). In fact, the opportunity to detect and punish cheaters in studies of multiple-trial Prisoner's Dilemma, Public Goods, and Trust Games was largely responsible for producing outcomes that deviate from standard game-theoretic predictions. So not only do we have a bias toward cooperating, we expect others to do so as well, and we will retaliate mightily if they don't.

Think about what that means for an economic theory based on self-interest: People will punish someone who defects in Prisoner's Dilemma, even though such a person is doing what game theory shows is the optimal strategy. Bystanders will even go so far as to pay a penalty so that they can have the opportunity to punish someone who defected in an observed Prisoner's Dilemma game, particularly if the person defected and their partner did not (Fehr & Fischbacher, 2004b). Plainly, people are motivated not only by economic self-interest in Prisoner's Dilemma games but also by norms of fairness and anticipated reciprocity.

Differences in Power and Status Influence How Fairly We Treat Others

Economists are even more puzzled by people's behavior in two other games, Dictator and Ultimatum. In one-shot Dictator, two strangers are given the opportunity to divide a sum of money. The catch is that the experimenter gives one of the parties (the dictator) total decision-making authority to divide the money any way he or she sees fit. The other party has no say in the matter. According to game theory, a rational dictator should keep all the money. Yet dictators typically offer the other party anywhere from 15% to 35% of the stake (Camerer, 2003). This is pure altruism; the game is one-shot, and these are strangers whose identities might be concealed from each other and even from the experimenter. Yet people will give up money they could just as easily take home without fear of repercussion.

In one-shot Ultimatum, the experimenter also assigns roles to two parties in order to divide a sum of money. But there is an important twist: The other party has a say in the matter. In Ultimatum, the person dividing the sum is the proposer who simply proposes how to split the money. The other party, the responder, can either accept or reject the offer. If the offer is accepted, the money is divided as proposed, and they both go home with money in their pockets. But if the responder rejects the money, the entire sum is forfeited; they both go home empty-handed.

If we define rationality as self-interest, then a rational proposer should offer slightly more than nothing, and a rational responder should take whatever is offered. After all, even one penny is better than no money at all, and both parties know that. But that's not what people do. Average offers in Ultimatum are a good deal higher than in Dictator – between 30% and 50% (Camerer, 2003). Offers less than 20% are typically rejected, meaning that people would rather no one get any money than accept an offer they believe to be too low. Again, people seem to approach these games with a *norm of self-interest* and *a norm of fairness* (Eckel & Grossman, 1995; Rabin, 1993).

Do people always operate according the principle of fairness meaning a fifty-fifty split? Not so much. Van Dijk and Vermunt (2000) had people play Dictator and Ultimatum games under conditions of symmetric (we both know everything we need to know) and asymmetric (one party knows something the other doesn't) information. The information manipulated was highly relevant, namely, that the dictators and proposers would be given double the value of each playing token, whereas their partners would receive only the stated value of the token. So if a token was marked "$1," dictators and proposers would receive $2 when they cashed in the token, whereas their partners would get $1. Dictators were unaffected by the information, making the same kinds of offers in both conditions. But proposers were very much affected; they made offers of nearly equal monetary value distribution when their partners knew about the true values of the tokens for each party, but they exploited their partner's ignorance in the asymmetric-information condition by making a seemingly fair offer to split the tokens in half. In reality, those seemingly equal distributions meant the proposer would end up with one-third more money. So dictators had more power and more information than their partners, and they behaved more fairly toward their partners. It was as though having that much of an advantage over their partners triggered the norm of fairness. But when proposers in the Ultimatum game had an informational advantage over their partners, they behaved selfishly. Or, as an economist would put it, when proposers had the advantage of asymmetrical information, they behaved strategically, like brokers who engage in insider trading or Enron executives who encouraged their employees to hold onto their stock options while they sold their own, knowing the company was in trouble and the stocks would be worthless very soon.

Using a different methodological approach, Fiddick and Cummins (2007) found even more intriguing results. People were asked to evaluate a carpooling arrangement in which one party agrees to pay for gasoline if the other party does all the driving. They were shown hypothetical ledgers showing gas payments that indicated varying degrees of compliance on the part of the gas-paying partner (from 100% compliance to as little as 25%). They were asked how willing they would be to continue the arrangement at each level of compliance and how fairly they thought the other person was treating them. The twist was that, in some scenarios, the two parties had equal status and power (both were employees), and in some they were of unequal status and power (one party was the other one's boss). People were far more tolerant of the employee cheating the boss than they were of the boss cheating the employee. This was true even when the employee was described as making a lot more money than the boss because the employee had a home-based computer business on the side. But one crucial factor had to be present for these asymmetries in tolerance and perceived fairness to occur: The employee had to work for that boss. If the parties were described as a boss and an employee from different companies who met through a classified ad, the effect disappeared; equivalent levels of intolerance for cheating were found regardless of employment status. So it seems it is asymmetries in the social relationship, and not asymmetries in costs and benefits, that underlie the effect.

Fiddick and Cummins referred to these results as the *noblesse oblige* effect. *Noblesse oblige* is a French term that can be roughly interpreted as "status and power entail obligation" or "with wealth, power, and prestige come responsibilities toward those less fortunate." In ethics, the term is sometimes used to describe a moral economy wherein privilege is balanced by duty toward those who lack such privilege. The generous behavior of the dictators in van Dijk and Vermunt's (2000) study may also be described this way. They were in total control of both assets and vital information, yet they did not take advantage of their partners. They behaved as though their advantages imbued them with pastoral responsibility toward their partners.

Contrast this with what happens when people size each other up purely on the basis of competitive performance. In a series of studies, Hoffman and colleagues (Hoffman, McCabe, Shachat, & Smith, 1994; Hoffman, McCabe, & Smith, 1996; Hoffman & Spitzer, 1985) manipulated relative standings in a competitive pre-game to see if these relative standings would affect the way people behaved in Dictator and Ultimatum games. Participants were required to complete a current-events test and were then ranked according to the number of correct answers they had. These rankings were posted where everyone could see the results. Then the experimenters paired people so that the pairs consisted of one high-ranking and one lower-ranking individual. The top-ranking person was paired with the person who ranked tenth, the second-highest ranked person was paired with the person who ranked eleventh, the third-highest was paired with the person ranked twelfth, and so on. The higher-ranking person in each pair was assigned the role of dictator in the Dictator game or the role of proposer in the Ultimatum game.

The results were striking. Dictators made significantly greater uneven distributions (favoring themselves) compared to a control condition where no pre-game was played. Proposers made significantly lower offers without raising the rejection rate (again, compared to a control condition). In other words, higher-ranking individuals thought they were entitled to more, and lower-ranking individuals thought they were entitled to less.

Neuroscience Makes It Clearer Why We Behave the Way We Do

So let's take stock: The results of experimental-economics studies show that decision makers are generally less selfish and less strategic than game theory predicts, and they value social factors such as reciprocity, fairness, and relative social status more than the theory predicts. Is this just because we keep making mistakes? Are we trying to behave in ways that an economist defines as rational, but we keep falling short of the mark because of human fallibility? Or are we wired this way? So far, the results of neuroscience studies suggests that the last of these is actually the case: We are wired this way.

As we will see in more detail in the next chapter the front part of the brain tracks probability information, whereas activity in deeper parts of the brain tracks the magnitude of reward or punishment. The reward or punishment could be monetary, or it could be social in nature. So what happens when we have people play Prisoner's Dilemma while undergoing fMRI imaging of their brains? Bottom line: Even when the same amount of money is gained or lost, reciprocated cooperation with another human leads to increased activation in the striatum (reward area) whereas unreciprocated cooperation shows a corresponding decrease in activation in this area (Sanfey, 2007). These results indicate that people find cooperation rewarding and lack of cooperation distressing in Prisoner's Dilemma games. Rilling and colleagues (2002) summarized it this way: Activation of the brain's reward circuitry positively reinforces cooperation, thereby motivating subjects to resist the temptation to selfishly defect.

In related work, neural reward circuitry was found to become active when people donated money to charity and when they observed money being donated to charity. But this was true only if the donations were voluntary; if the donations were made because they were required by someone's job or other involuntary means, these reward circuits did not light up. This means that we find it rewarding to behave altruistically, not just cooperatively.

We know that people will go out of their way to punish defectors in Prisoner's Dilemma, even when they are only observing the game rather than playing it themselves, and even when it costs them money to do so. It turns out that when players are given the option to punish defectors, activation in the brain's reward circuitry occurs, even when the person is losing money to punish the other party. This means that people find it rewarding to punish defectors. Given that Public Goods games have the same kind of mathematical and social structure as Prisoner's Dilemma, it should come as no surprise that these games yield the same neural-imaging results: Reward-related brain areas were activated when free riders were punished (de Quervain et al., 2004).

In the Trustee game, activity in reward pathways of the brain was greatest when the investor repaid generosity with generosity and most subdued when the investor repaid generosity with stinginess (King-Casas and colleagues, 2005). Even more surprisingly, the amount of money investors were willing to fork over could be manipulated chemically using a substance called oxytocin. Oxytocin is sometimes referred to as the "bonding" hormone; it seems to facilitate social bonding and trust. It is a hormone secreted from the posterior lobe of the pituitary gland. It initiates labor in pregnant women and facilitates production of breast milk. This means that both baby and mother are flooded with a feel-good social-bonding hormone during birth and nursing. Oxytocin levels also rise in men and women during sex, again facilitating emotional bonding. In a study by Kosfeld and colleagues (2005), investors and trustees were given oxytocin or a placebo before playing the Trustee game. Oxytocin increased the willingness of the investors to trust – they forked over more money. But it had no impact on the trustees. They behaved the same as trustees always do in these studies, returning slightly less than was invested. So oxytocin made people more willing to trust in this study but not more generous. Similar results were found in the Ultimatum game: Intranasal oxytocin increased generosity by 80% but had no effect in the Dictator game (Zak, Stanton & Ahmadi, 2007), again indicating that oxytocin may make us more trusting but not necessarily more generous. And how do we feel about stingy and generous offers? Stingy offers in Ultimatum games activate brain areas (anterior insula) associated with feelings of disgust (Sanfey and colleagues, 2003). But this is true only if a person is playing with another person; if the other party is a computer, no changes occur in those brain areas.

The results of decision-neuroscience studies like these plainly show that the impact of social aspects of these games cannot be overestimated. People behave as though they are wired to expect long-term reciprocal relationships. Reward circuits become active when we behave cooperatively and generously, and disgust circuits become active when we behave selfishly. The outcome of all this wiring seems to be an attempt to achieve the following social goals: Increase the likelihood that inequity is avoided, foster mutual reciprocity, and encourage punishment of those seeking to take advantage of others. In repeated games – repeated transactions among individuals who will remember one another and their transaction history – reputation becomes exceedingly important. No one wants to engage in transactions with individuals who have a reputation for stinginess and selfishness – not even those who may be planning to behave that way themselves. Whether you're selfish or generous, transacting with a cooperator is always the better bet.

In fact, this point is obvious even to infants. In a set of studies (Hamlin and colleagues, 2007), six-month-old infants watched as a red disc struggled to move up a steep incline. In one condition, a yellow triangle came racing along and pushed the red disc to the top of the incline. In another, a blue square came racing along and pushed the red disc down to the bottom of the incline. In other conditions, a third object either did nothing or was inanimate and could not move. After viewing the show, the infants were shown the three objects and allowed to select which one they wanted to play with. The infants overwhelmingly preferred helpers (cooperators) over neutral parties and neutral parties over those who hindered. The authors concluded that even preverbal infants assess individuals on the basis of their behavior toward others and, moreover, that this kind of social evaluation is a biological adaptation.

The Evolution of Cooperation

Decision neuroscience shows us that we seemed to be wired for cooperation, and developmental science seems to show that we are more than wired this way – we are born this way. But inquiring minds may still feel unsatisfied. We want to know *why* we are born this way. The best answer to this question has come from the field of evolutionary biology.

As Axelrod and Hamilton pointed out in their classic 1981 paper on the game-theory computer tournament, "The theory of evolution is based on the struggle for life and the survival of the fittest. Yet cooperation is common between members of the same species and even between members of different species." The phenomenon of cooperation is difficult to reconcile with a selfish-gene view of evolution – that is, the view that the genes whose consequences serve their own implicit interests to continue being replicated are the ones that are passed on (Dawkins, 1976). When we act selfishly (by, for instance, keeping our food to ourselves), we enhance our chances of survival. When we enhance our chances of survival, we live longer lives. When we live longer lives, we have more opportunities to reproduce. Hence, our genes are more likely to remain – and spread – through the population. So it is easy to see how genes that support selfish behavior can flourish. So how do genes that support non-selfish behavior – cooperation – flourish?

Hamilton suggested one answer: Kin selection. Kin selection can be explained simply this way: If you share a lot of genes with another individual, then as long as the cost of helping them is less than the benefits they receive from your assistance, the genes you share will benefit and flourish. Suppose you have a lot of money, and your sister is struggling to take care of her son. If you share a little of your money with her so she can take care of your nephew, you have benefited your shared genes and incurred very little cost to yourself. So helping your genetic relatives can still be accommodated by a selfish-gene view.

But how do we explain cooperation among non-relatives? This happens quite frequently in nature. Consider the phenomenon of cleaner-fish: They nibble parasites and dead tissues off the mouths of other fishes and then exit through the mouths and gills. Cleaners are allowed to approach with little danger of being eaten. So the host fish benefits from a cleaning, and the cleaner-fish benefits with a nice meal. A little more sophisticated case is that of vampire bats. These are essentially large mosquitoes in that they feed on the blood of other animals. They can share blood with other bats but usually share with only those that have shared blood with them in the past (Wilkinson, 1984). In the world of primates, reciprocation occurs frequently among non-relatives. Vervet monkeys are more likely to respond to calls for help from non-kin in agonistic encounters if the caller has groomed them recently, and they also form the strongest alliances with individuals who groom them most often (Cheney & Seyfarth, 1992). And like humans in a Public Goods experiment, chimpanzees retaliate against individuals who don't share food either by direct aggression when they request food (de Waal, 1989) or with misinformation or lack of information about the location of food (Woodruff & Premack, 1979).

Notice that, in each of these cases, both parties in the cooperative venture benefit, so both sets of genes benefit. Why would selection favor benefiting someone else's genes, especially when in each transaction, the cooperator could do better by defecting? Why doesn't the host fish just eat the cleaner-fish after the latter has done its work? Why do the vampire bats and primates share, and why do they tolerate a cost to themselves to punish those who don't?

In 1971, Robert Trivers offered one very influential solution to this puzzle. He showed that altruism can be selected for when it benefits the organism performing the altruistic act as well as the recipient. He called this kind of cooperation *reciprocal altruism*, and he described it as "… each partner helping the other while he helps himself." The problem with reciprocal altruism is that an individual can benefit from cooperating, but he or she can usually do better by exploiting the cooperative efforts of others. Trivers showed that selection will discriminate against cheaters if cheating later adversely affects the cheater's life to an extent that outweighs the benefit of not reciprocating. When will this happen? When a cooperator responds to cheating by excluding cheaters from future transactions. Under these conditions, reciprocal altruism is an evolutionarily stable strategy (ESS). The concept of an ESS was introduced by John Maynard Smith in 1972. A strategy is called evolutionarily stable if a population of individuals playing this strategy is able to outperform and eliminate a small subpopulation playing a different strategy.

Box 2.2. The Mathematics of Reciprocal Altruism

Consider two populations, altruists (*A*) and non-altruists (*N*)
A performs altruist act when cost to self is less than the benefit to other

Cost (df) reduction in reproduction of genes
Benefit (df) increase in reproduction of genes

Assume altruistic behavior is controlled by an allele at a particular locus (a_2), and there is only one other alternative allele possible for that locus (a_1) that leads to non-altruistic behavior. How should altruistic behavior be dispensed?

Random Dispensation of Altruism

Three possible genotypes: a_1a_1, a_1a_2, a_2a_2.

a_1a_1 (Non-Altruist) benefits by $(1/N)\Sigma b_i$, where b is benefit to recipient.
a_2a_2 (Altruist) has a net benefit of $(1/N)\Sigma b_i - (1/N)\Sigma c_j$, where c is the cost to the a_2a_2 actor.

Result: $(1/N)\Sigma c_j < 0$, so a_1 will replace a_2 in the population.

Nonrandom Dispensation of Altruism to *Kin* Only

Altruism will spread if benefits to recipient greatly outweigh costs to actor (a_2 will replace a_1) (Hamilton, 1964) $r > c/b$

Nonrandom Dispensation to *Reciprocators* Only

Altruism will be selected if the net benefit accruing to a_2a_2 altruist exceeds that accruing to an a_1a_1 non-altruist

$(1/p^2)(\Sigma b_k - \Sigma c_j) > (1/q^2)\Sigma b_m$ where

b_k is the benefit to a_2a_2 recipient
c_j is the cost to the a_2a_2 actor

b_m is the benefit to the a_1a_1 recipient
p is the frequency of the a_2 allele
q is the frequency of the a_1 allele

Necessary Condition: Σb_m must remain small

When that will happen: When the altruist responds to cheating by curtailing future transactions with that individual.

When modeling reciprocal altruism as in Prisoner's Dilemma, the payoff matrix has these constraints: $T > R > P > S$ and $R > (S+T)/2$. This means that the temptation to defect must be greater than the reward for cooperating, which in turn must be greater than the punishment for not cooperating. The sucker's payoff has the lowest payoff – you get stuck holding the bag if you cooperate with a defector. Finally, the reward should be greater than half the sum of the sucker's payoff and the temptation to defect. This kind of matrix best captures the scenario underlying the Prisoner's Dilemma.

Several researchers tried it. They modeled reciprocal altruism as Prisoner's Dilemma in a way that satisfied Trivers' conditions. In one-shot Prisoner's Dilemma, defection is an ESS. But in repeated games, cooperation is an ESS when participants can recognize each other and cheaters are excluded from future cooperative ventures.

So let's think about that: If you've never seen the other party before, have no background knowledge or shared history with that person, and neither of you will see the other again, then the strategy with the greatest payoff is take the money and run. But if you know the other party through reputation or a shared history, and you both know you're going to run into each other frequently in the future, cooperation wins out in the long run. But only if you stay away from cheaters – those who take the money and run. If you continue to transact with cheaters, then, in every simulation, altruism goes extinct (Sober & Sloan-Wilson, 1999). So cheaters must be excluded from future transactions; otherwise, cooperation disappears from the population.

Trivers also listed the following characteristics that favor the evolution of cooperation: Long lifespans of individuals in the population, low dispersal rate (people stick around), interdependence among the members, high degree of parental care (which aids kin selection and altruism), needing aid in defense and combat, and the absence of a rigid linear-dominance hierarchy. As it turns out, this description exactly fits the hunter-gatherer lifestyle of early humans. So, according to Trivers, our species evolved under conditions that favor selection for cooperation. Individuals in ancestral populations who behaved this way survived longer and reproduced more than those who didn't behave this way, so their altruism genes became more common over many generations. That is why we seem to be born this way. That is why we enter these social or monetary transactions with a strong bias toward cooperating. That is why we expect reciprocity and why we will go out of our way to punish those who fail to reciprocate. We are wired to preserve cooperation because that is what allowed us to survive in the long run.

THREE

Rational Choice

CHOOSING WHAT IS MOST LIKELY TO GIVE YOU WHAT YOU WANT

In his 2002 book, *Calculated Risks*, Gerd Gigerenzer reports the case of a doctor who convinced 90 "high-risk" women without cancer to sacrifice their breasts "in a heroic exchange for the certainty of saving their lives and protecting their loved ones from suffering and loss." But as Gigerenzer points out, if the doctor had done the calculations correctly, he would have found that the vast majority of these women (84 out of 90, to be exact) were not expected to develop breast cancer at all. Now here is a question: Recently, the American Medical Association changed the rules concerning screenings for breast cancer and prostate cancer, claiming that frequent testing returned too many false positives (test results saying cancer was present when it was not). They recommended less frequent testing. And the outcry among patients was deafening. Television pundits claimed that the decision was driven solely by a desire to reduce medical costs and that this meant the medical profession was willing to sacrifice lives in an effort to save money. Was this decision medically motivated, financially motivated, or (as in the preceding example of the "heroic" doctor) just another case of misguided reasoning? This chapter will help you decide.

How the "Big Boys" Think about Decision Making

The concept of a rational agent is the cornerstone of classic decision theory. But what the average person means by "rational" is not necessarily the same thing an economist means. You may describe people as rational if their thinking is calm, logical, and coherent. To an economist, rational agents are first and foremost self-interested – that is, they compare choices and select those that maximize their own benefits and minimize their own costs.

Rational choice theory can be summarized this way: We make decisions by determining how likely something is to happen, judging the value of the outcome, and then multiplying the two (Edwards, 1955). The optimal decision is the one with the largest product – that is, the decision most likely to give us what we want most.

Notice that there are two parts to the decision-making process. The first is determining the likelihood of the desired outcome. The second is deciding what it's worth to you. Simple example: Should you try for a date with Brad Pitt, the nice, plain-looking guy next door, or the boring guy your mother keeps trying to get you to date? Brad's the best looking and lives a pretty exciting life, so a date with him would be more desirable than a date with the plain guy next door. A date with the boring guy is least desirable. But getting a date with Brad isn't very likely, getting a date with the guy next door is more likely, and getting a date with the boring guy who's been hounding your mom to fix him up with you is a sure bet. All told, opting for a date with the guy next door is the best choice: It's more likely to happen than a date with Brad (whom you find more desirable), and a lot more desirable than a date with mom's guy (which is a sure thing). So you've just made a rational choice.

The desirability of a choice is called its utility. It is a measure of satisfaction. Winning $1,000 is obviously more satisfying than winning $100, so we would prefer winning $1,000 to winning $100. Rational choice theory makes two assumptions about individuals' preferences for actions. The first is *completeness* – all actions can be ranked in an order of preference (indifference between two or more is possible). The second is *transitivity* – if choice 1 is preferred to choice 2, and choice 2 is preferred to choice 3, then choice 1 is preferred to choice 3. Together, the assumptions of completeness and transitivity mean that individuals can rank choices in terms of preferences and that their preferences are consistent. So using our example, your preferences regarding your potential date could be ordered like this: Brad > guy next door > boring guy. If I gave you a choice between Brad and the guy next door, you should choose Brad. If I gave you a choice between the guy next door and boring guy, you should choose the guy next door. And if I gave you a choice between Brad and boring guy, you should choose Brad. If you don't, then your choices violate the norm of transitivity, and I have no idea how to predict what you are going to choose.

The concept of utility is usually pretty clear to people; it's whatever makes you feel happier or more satisfied. But the second part of rational choice – calculating probability – is usually the fly in the ointment. In our dating example, we guessed the probability of getting dates with each of the three guys. But what if you are asked to make a decision about, for example, the probability that you actually have cancer given that you've gotten a positive mammogram result or a positive test result for a prostate-specific antigen?

Let's start with breast cancer. To answer this question, we need some information that can be obtained from the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Cancer Statistics report (www.seer.cancer.gov). From this report, we can see that 12% of women born today will be diagnosed with breast cancer at some time during their lifetime. This is called the lifetime risk of developing cancer, and it is based on the cancer incidence rate – that is, the number of new breast cancers that occurred in the United States during a year. It tells you the number of new cancers per 100,000 people. This is different from prevalence, the number of people in a specific population who have a certain type of cancer at a specific point in time. So incidence tells you the estimated number of new cases of a cancer, whereas prevalence tells you the number of all cases.

But mammograms are not perfect predictors of cancer. The machine takes the X-ray, but a trained radiologist must review the X-ray and make a decision about whether or not cancer is present. The sensitivity of the test tells you the percentage of cancers that give true positive mammogram results. Their specificity tells you the percentage of healthy cases that give true negative results. But radiologists can make mistakes. A false positive means the radiologist decided you have cancer when you don't. A false negative means that the radiologist decided you don't have cancer, but you do.

Table 3.1. *Statistics for Breast Cancer and Mammograms According to Patients' Age*

Breast Cancer			Mammograms			
Age	Incidence (%)	Prevalence (%)	Age	Positive (%)	Sensitivity (%)	Specificity (%)
30–39	.43	.17	30–39	Not available		
40–49	1.45	.89	40–44	18.2	88.2	81.9
50–59	2.38	2.18	50–54	22.1	90.9	78.3
60–69	3.45	3.75	60–64	19.3	88.8	81.5

All of these values vary as a function of your age. Table 3.1 presents the rates for breast cancer in American women according to the SEER report and information on mammograms from the Breast Cancer Surveillance Consortium (www.breastscreening.cancer.gov).

Notice that both incidence and prevalence of breast cancer increases with age. This means that your risk of developing breast cancer increases with age. Other factors may also put you at greater risk, such as family history of breast cancer or environmental factors. But these numbers show that age is a major factor. Notice also that the percentage of positive mammograms increases as well, whereas the percentage of cancers correctly diagnosed (sensitivity) and the percentage of healthy cases accurately diagnosed as healthy (specificity) don't change that much.

Suppose you've gotten your first mammogram, and the results are positive for breast cancer. What is the probability that you actually have breast cancer? Using these numbers, the probability of a 40-year-old having breast cancer given a positive mammogram is about 5%; for a 50-year-old, the probability is 9%.

Did these numbers surprise you? Did you think the probabilities would be at least ten times higher than that? If you did, you're not alone. Your physician might have made the same error: When physicians were given problems like these to solve, the results were dismal (Eddy, 1982). Their estimates were off by a factor of 10 – that is, if the actual probability of having cancer given a positive test was 9%, the average estimate given by the physicians was 90%! So it might be a good idea for you, as an informed patient, to know how to calculate this.

To make this judgment, you will need to make use of Bayes Theorem. Thomas Bayes (1702–1761) was an English country clergyman, amateur mathematician, and inveterate gambler. He was also an inveterate gambler who was keenly interested in how best to figure the chances of drawing a winning hand in card games, throwing the right combination of numbers with a pair of dice, or picking the winner in a horse race. In his *Essay Towards Solving a Problem in the Doctrine of Chances* (partially rewritten by Richard Price and published in 1763), he proposed a way of making decisions based on calculating the probabilities.

Bayes Rule is considered the optimal statistical model for making decisions about risk. I should mention that risk is distinct from uncertainty. When calculating risk, you base your calculations on probabilities that are known, as in breast cancer. If you don't know the probabilities associated with various events or choices, then you can't use Bayes Rule. In "How to be more Bayesian" we will calculate it using frequencies, which is much easier. You can skip to that section now, if you like.

Box 3.1. Bayes and a Deck of Cards

Let's translate all of this talk of probabilities into a domain most of us will find familiar. Suppose I pull a card from a normal deck of playing cards. What is the probability that the card is king? That's easy; there are only 4 kings in the deck, and there are 52 cards in the deck. So the probability that the card is a king is 4 out of 52. This is the prior probability of pulling a king from the deck. We will symbolize it as P(King).

Suppose I tell you that the card is indeed a king, but now I ask you this: What is the probability that it is the king of diamonds? Well, there are 4 kings, and 1 of them is the king of diamonds. So the probability that the card is the king of diamonds is 1 out of 4. This is the posterior probability – the conditional probability that the card is the king of diamonds given that I know the card is a king. We will symbolize it as P(Diamond|King).

Now let's gamble – that is, let's combine utility and probability. If you win the bet, you win $10. If I pull a card from this deck, which bet is best?

A. The card is a king.
B. The card is the king of diamonds.

(*continued*)

This is a no-brainer if you think in terms of odds or probabilities. There are 4 kings in the deck. And there is only 1 king of diamonds. So the odds are:

A. P(King) = 4 out of 52
B. P(King of Diamonds) = 1 out of 52

Plainly, bet A is the best one. This example shows you a very important rule concerning probabilities – the conjunction rule. This rule states that the conjunction of two events is always less probable than the probability of either event occurring alone. We can symbolize this as $P(A \& B) \leq P(A)$. Using our example, $P(\text{King \& Diamond}) \leq P(\text{King})$

Let's try another one: Suppose I pull a card from the deck, and I tell you that it is a diamond. Now, I offer you the following bets:

C. The card is a face card
D. The card is the king of diamonds

Again, this is a no-brainer. There are 13 diamonds in the deck. There are 3 kinds of face cards (king, queen, jack) so the odds that the card is a face card given that you know it is a diamond is 3 out of 13. There are 13 diamond cards in the deck, so the odds that the card is the king of diamonds given that you know the card is a diamond is 1 out of 13 because there is only one king of diamonds.

C. P(Face Card|Diamond) = 3 out of 13
D. P(King|Diamond) = 1 out of 13

Bet C is the best bet.

Notice that what you have done in this one is update your beliefs or expectations based on new information. If you didn't know that the card was a diamond, you would just bet using the prior probabilities A and B. But you have updated information; you know that the card is a diamond. So you bet by choosing between the posterior probabilities C and D.

What I've introduced to you intuitively is Bayes' Rule – a normative model for updating probabilities about the truth of hypotheses based on new observations or experiments.

Let's apply it to card cases C and D:

For case C, we need to calculate the posterior probability of P(Face Card|Diamond). To do this we need the prior probabilities of face card and diamond. That's easy. There are 12 face cards in the deck (king, queen, and jack for each of the four suits, diamonds, hearts, clubs, and spades), so the prior probability of pulling a face card is 12/52 = .23. There are 13 diamonds in the deck, so the prior probability of pulling a diamond is 13/52 = .25. Next, we need to figure the likelihood, which is the conditional probability that the card is a diamond given that we know the card is a face card. There are 12 face cards in the deck, and 3 of them are diamonds, so P(Diamond|Face Card) is 3/12 = .25. Now we are ready to calculate the posterior probability for the bet C.

$$\begin{aligned} \text{P(Face Card}|\text{Diamond)} &= \text{P(Diamond}|\text{Face Card)[P(Face Card)/} \\ &\quad\ \text{P(Diamond)]} \\ &= .25(.23/.25) \\ &= .23 \end{aligned}$$

We calculated the odds for case C, and they turned out to be 3/13, which is the same as .23.

For case D, we need the prior probability of pulling a king (4/52 = .08), the prior probability of pulling a diamond (.25), and the likelihood of pulling a diamond given that the card is a king (1/4 = .25). So the posterior probability of pulling a king given that the card is a diamond is

$$\begin{aligned} \text{P(King}|\text{Diamond)} &= \text{P(Diamond}|\text{King)[P(King)/P(Diamond)]} \\ &= .25(.08/.25) \\ &= .08 \end{aligned}$$

We calculated the odds for case D to be 1/13, which is the same as .08. So when you thought about the frequencies of different types of cards in a deck of cards, you made decisions that followed Bayes Rule – even if that seemed a lot easier than converting everything to probabilities and plugging them into Bayes formula.

Bayes Rule simply states that the probability of a hypothesis given the data (the posterior) is proportional to the product of the likelihood times the prior probability. The likelihood shows the effect of the data. Here is the equation:

$$\text{Posterior Probability (A} \mid \text{B)} = \frac{\text{Likelihood (B} \mid \text{A)} * \text{Prior Probability (A)}}{\text{Prior Probability (B)}}$$

Now let's use Bayes Rule to find out what the chances are that you have breast cancer given that you are 40 years old and have gotten a positive result on your first mammogram. The prevalence of breast cancer in your age group is 1%, so the base rate for the probability of breast cancer in your age group is .01. About 18% of mammograms are positive in this age group, so the base rate probability of getting a positive test is .18. Radiologists correctly diagnose cancer when it is present 88% of the time in this age group, so the probability that you will get a positive test result from the radiologist if you have breast cancer is .88. Now we just plug all the numbers in:

$$\begin{aligned} P(\text{Cancer}|\text{PosTest}) &= P(\text{Pos Test}|\text{Cancer})\ [P(\text{Cancer})/P(\text{Pos Test})] \\ &= .88\,(.01/.18) \\ &= .05 \end{aligned}$$

In other words, at age 40, your chances of actually having breast cancer if your mammogram is positive is 5%. Suppose you are 50 years old. Then the calculation becomes .91(.0218/.22) = 9%.

Keep in mind that the calculation always takes into account the actual prevalence of breast cancer in your age group. As the prevalence rate shows (e.g., 2.18%), breast cancer is extremely rare in each age category. The lifetime risk of developing cancer – the probability of developing it over your entire lifespan – is 12%, or 1 in 8. But only 1% of women in the age group 40 to 49 actually have breast cancer (1 out of 112). By age 50, this has increased to a little more than 2% (1 out of 46).

Now the controversy: Should women have mammograms annually, and at what age should they start? Your gut reaction may be something like "the sooner the better." But here is the problem: Mammograms are not like photographs. You can't just look at mammograms and see cancer. The films must be viewed and interpreted by trained radiologists, and there is a nontrivial error rate associated with those interpretations. According to the 2009 Breast Cancer Consortium report, radiologists incorrectly diagnose breast cancer in 40-year-old women who are getting their first mammograms about 8% of the time. This means that about 1 out of 12 of these women may be told she has cancer when she in fact doesn't. The false-positive rate for 50-year-old women is 12% (about 1 out of 8 women)! So every time you have a mammogram, you run the risk of getting a false positive. In fact, your chances of getting a false positive increases significantly with the number of tests you have over your lifetime. For example, a woman who has 10 mammograms has a 49% chance of being called back at least once because of a false positive (Christiansen et al., 2000). If you start having mammograms at age 40, this means you have almost a 50% chance of being told you have breast cancer when you don't by the time you are 50. Because of this, the U.S. Preventive Services Task Force issued new guidelines in 2009 recommending that screening should begin at age 50 rather than 40 and that women should have mammograms every two years rather than every year. This was not a popular recommendation, and outraged women and their doctors protested. The primary difficulty seems to be convincing people that a positive mammogram is not sure evidence of cancer. (For more on how to interpret medical-screening results, see Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2008.)

In 2008, the guidelines regarding prostate cancer screening also changed. The most prevalent test used is called the prostate-specific antigen (PSA) test. PSA is produced by both normal and cancerous glands. As men age, both benign prostate conditions and prostate cancer become more common. As a result, interpreting a rise in PSA in terms of benign or cancerous conditions is not an easy task. Previously, men were advised to begin screenings at age 40. But then the results of the *U.S. Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial* were published. The researchers in this long-term study had randomly assigned more than 76,600 men to two groups: those who received "usual care" and those who had annual PSA tests for six years and digital rectal examinations every year for four years. The researchers found little difference in prostate-cancer death rates between the two groups at seven years or again at ten years of follow-up.

In his book, *Statistics for Management and Economics*, Gerald Keller (2001) reports data on PSA and prostate cancer and uses this information to calculate the posterior probability that a 40- to 50-year-old man has prostate cancer given that he has had a positive PSA test (PSA score of 4.1 or higher). This probability turned out to be .05! Based on results like these, the American Cancer Society recommends that

> men make an informed decision with their doctor about whether to be tested for prostate cancer. Research has not yet proven that the potential benefits of testing outweigh the harms of testing and treatment. Starting at age 50, talk to your doctor about the pros and cons of testing so you can decide if testing is the right choice for you. If you are African American or have a father or brother who had prostate cancer before age 65, you should have this talk with your doctor starting at age 45.

In October 2011, The U.S. Preventive Services Task Force went even further in recommending that men be told the pros and cons of testing so that they can decide for themselves. The Task Force examined all the evidence and found little, if any, reduction in deaths from routine PSA screening. What happens instead is that too many nonfatal tumors are discovered and treated aggressively, yielding serious side effects from treatment that is practically unnecessary.

This means that the doctor and patient must now make a Bayesian decision based on probability information. But, as we've seen, even doctors frequently make mistakes when basing their decisions on probabilistic information. According to Gerd Gigerenzer, this is because expressing information in terms of probabilities and percentages has only been around since the 1700s, but our cognitive systems evolved over millions of years to process information in frequency format. In fact, you can visualize frequencies without even counting (Zacks & Hasher, 2002). It's called subitizing, and we do it automatically. For example, when you look around your environment, you might see four people wearing red shirts and two people wearing blue shirts. You don't see 66% of people wearing red shirts and 24% wearing blue ones. You don't see a .66 probability of red shirts and a probability of .34 of blue shirts. You can express your frequency information that way, but what your perceptual/cognitive system registers is quantities or frequencies. It "understands" frequencies.

Table 3.2 shows how the breast-cancer data look if we express prevalence in terms of frequencies (or odds).

Table 3.2. *Prevalence of Breast Cancer According to Age of Patient*

Age	Prevalence	
30–39	1.7 out of 1,000	1 out of 588
40–49	8.9 out of 1,000	1 out of 112
50–59	21.8 out of 1,000	1 out of 46
60–69	37.5 out of 1,000	1 out of 26

How to Be More Bayesian

Now, what if we were to translate the probability information in the breast-cancer problem into frequency information in order to figure out what a positive mammogram means? Would this empower us to make wiser decisions? Let's try it:

> *Imagine 1,000 women at age 40 who are undergoing their first mammogram screening for breast cancer.*
>
> *Nine of them will have breast cancer and will get a positive mammogram.*
>
> *One hundred seventy-eight of them will not have cancer and will get a positive mammogram.*
>
> *What are the chances that a woman who has a positive mammogram actually has cancer?*

That's easy: 9 out of 178 (which is equivalent to 5%, or 1 out of 20). To make it even clearer, a positive mammogram means you have a 1-in-20 chance of actually having breast cancer and 19 out of 20 chances that you don't. When decision-making information is presented to people in this kind of frequency format, accuracy rates increase dramatically; this is true even for physicians (Chase, Hertwig, & Gigerenzer, 1998; Gigerenzer & Hoffrage, 1995). As these authors point out, expressing risk this way makes the calculations much easier for us, hence we are less prone to errors.

Just to be clear, here is where the 9 and 178 came from: We know that the cancer rate for this age group is 1%, so this means 10 women out of the 1,000 will be expected to have breast cancer and 990 will not. But radiologists correctly diagnose breast cancer when it is present in this age group about 88% of the time. 10 × .88 = 9, so nine of these women will have breast cancer and will be properly diagnosed, but one will receive a false negative – she has breast cancer, but the x-ray will be interpreted to mean she does not. We also know that radiologists incorrectly diagnose breast cancer when it isn't present in this age group about 18% of the time. So out of the 990 cancer-free women, 178 will get a positive mammogram (990 × .18 = 178). Those 178 might then receive unnecessary treatment, not to mention a good deal of anxiety over the diagnosis. Compare this with what happens at age 50:

> *Imagine 1,000 women at age 50 who are having their first mammogram screening for breast cancer.*
> *Twenty of them will have breast cancer and will get a positive mammogram.*
> *Two hundred fifteen of them will not have cancer and will get a positive mammogram.*
> *What are the chances that a woman who has a positive mammogram actually has cancer?*

That's easy: 20 out of 235 positive tests (about 9%, or 1 out of 11). Again, notice that 215 women will receive false-positive results, and they might receive unnecessary treatment. But 1 of out 11 is more worrisome than 1 out of 20, so that is why the recommendation is to start mammogram screening at age 50.

Here are more examples of the dramatic improvement in our ability to think like Bayesians when the information is given to us in frequency format. These examples are also taken from Gigerenzer's book *Calculated Risks*:

> QUESTION*: If men with high cholesterol have a 50% higher risk of heart attack than men with normal cholesterol, should you panic if your cholesterol level is high?*
>
> ANSWER: 6 out of 100 men with high cholesterol will have a heart attack in 10 years, versus 4 out of 100 for men with normal levels. In absolute terms, the increased risk is only 2 out of 100 – or 2%. Look at it this way: Even in the high-cholesterol category, 94% of the men won't have heart attacks.

> QUESTION: *HIV tests are 99.9 percent accurate. You test positive for HIV, although you have no known risk factors. What is the likelihood that you have AIDS, if 0.01 percent of men with no known risk behavior are infected?*
>
> ANSWER: Fifty-fifty. Take 10,000 men with no known risk factors. One of these men has AIDS; he will almost certainly test positive. Of the remaining 9,999 men, 1 will also test positive. Thus, the likelihood that you have AIDS given a positive test is 1 out of 2. A positive AIDS test, although a cause for concern, is far from a death sentence.
>
> QUESTION: *In his argument to the court to exclude evidence that O.J. Simpson had battered his wife, Alan Dershowitz successfully argued that the evidence was irrelevant because, although there were 2.5–4 million incidents of abuse of domestic partners, there were only 1,432 homicides. Thus, he argued, "an infinitesimal percentage – certainly fewer than 1 of 2,500 – of men who slap or beat their domestic partners go on to murder them." Was he right?*
>
> ANSWER: Think of 100,000 battered women. Forty will be murdered this year by their partners. Five will be murdered by someone else. Thus, 40/45 murdered and battered women will be killed by their batterers – in only 1/9 cases is the murderer someone other than the batterer.

When We Are Not Bayesian

To bring this point home even further, let's see what we do when we are required to make decisions based on probability information. This is a study by Kahneman and Tversky (1973). Try your hand at answering the questions.

> *A panel of psychologists have interviewed and administered personality tests to 30 engineers and 70 lawyers, all successful in their respective fields. On the basis of this information, thumbnail descriptions of the 30 engineers and 70 lawyers have been written. You will find on your forms five descriptions, chosen at random from the 100 available descriptions. For each description, please indicate your probability that the person described is an engineer, on a scale from 0 to 100.*
>
> *The same task has been performed by a panel of experts, who were highly accurate in assigning probabilities to the various descriptions. You will be*

> *paid a bonus to the extent that your estimates come close to those of the expert panel.*
>
> 1. *Jack is a 45-year-old man. He shows a good deal of interest in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing, and logic puzzles. The probability that Jack is one of the 30 engineers in the sample of 100 is _____ %.*
> 2. *Tom is 34 years old. He is intelligent, but unimaginative, compulsive, and generally lifeless. In school, he was strong in mathematics but weak in social studies and humanities. The probability that Tom is one of the 30 engineers in the sample of 100 is _____ %.*
> 3. *Suppose now that you are given no information whatsoever about an individual chosen at random from the sample. The probability that this man is one of the 30 engineers in the sample of 100 is _____ %.*

If you are like the majority of people in this study, your percentage for Jack was less than 30%, your percentage for Tom was greater than 30%, and your percentage for the third statement was 50%. But this problem stated that there were 30 engineers and 70 lawyers. The probability that any of them is an engineer is 30%. Tversky and Kahneman showed that when making decisions like this, people tend to ignore prior probabilities (the base rates for the different cases). Suppose this were a deck of cards, and they were asked about the probability of pulling a diamond from the deck. If they answered the way they did in this study, they would have said that the probability of diamonds was less than 25% for No. 1, greater than 25% for No. 2, and 50% for No. 3!

We don't even need to use numerical information to demonstrate departures from Bayesian reasoning. Do you remember the conjunction rule – that the probability of two events occurring together is less than the probability of either one of them occurring alone? In a number of studies, people were given a list of statements and were required to rank them in order of probability, starting with the most probable proposition and ending with the least probable proposition. Here is one example (Tverksy & Kahneman, 1983):

> *Tom is 34 years old. He is intelligent, but unimaginative, compulsive, and generally lifeless. In school, he was strong in mathematics but weak in social studies and humanities.*

A: Tom is an accountant.
B: Tom is a physician who plays poker for a hobby.
C: Tom plays jazz for a hobby.
D: Tom is an architect.
E: Tom is an accountant who plays jazz for a hobby.
F: Tom climbs mountains for a hobby.

The vast majority of people indicated that Tom was more likely to be an accountant who played jazz than a jazz player. That is, they believed the conjunction of "accountant and jazz player" was more probable than "jazz player." This is called the conjunction fallacy – when people judge a conjunction of events as more probable than one of them. We can symbolize this as P(A&J) > P(J).

Here's another:

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice and also participated in anti-nuclear demonstrations.

Rank these eight propositions by probability, starting with the most probable and ending with the least probable.

(a) Linda is a teacher in elementary school.
(b) Linda works in a bookstore and takes yoga classes.
(c) Linda is active in the feminist movement.
(d) Linda is a psychiatric social worker.
(e) Linda is a member of the League of Women Voters.
(f) Linda is a bank teller.
(g) Linda is an insurance salesperson.
(h) Linda is a bank teller and is active in the feminist movement.

Tversky and Kahneman (1983) gave this problem to undergraduates at the University of British Columbia; 92% ranked *h* higher than *f*. They also gave the problem to graduate students in the decision-science program at Stanford University Business School, all of whom had taken advanced courses in probability and statistics. It didn't matter; they, too, fell prey to the conjunction fallacy, with 83% ranking *h* higher than *f*.

Why do people fail so dismally at these tasks? For three reasons. First, these problems ask people to make probability judgments rather than frequency judgments, and as we've seen, most people have difficulty thinking in terms of probabilities. Second, despite asking about probabilities, these descriptions did not make clear that the cards were drawn randomly. When you explicitly tell people that, they pay more attention to base rates, and their performance improves (Gigerenzer, Hell, & Blank, 1988). Finally, and perhaps most importantly, these problems pit probability information against similarity judgments. People are very good at classifying items according to similarity to a prototype. In fact, we do it automatically. Although these problems ask people to make a probability judgment, the way they are written implicitly invites you to make a classification judgment.

Tversky and Kahneman called this the representiveness heuristic: The subjective probability of an event is determined by the degree to which it is similar in essential characteristics to its parent population. Sample size and prior probabilities are completely ignored. In the first problem, you automatically found yourself comparing the descriptions to a prototypical engineer, and you answered according to the similarity of the description to a prototypical engineer. In the second, you answered according to the similarity of the descriptions to a prototypical mathematician. The third invited you to think about a prototypical bank teller.

What if we were to ask people for a frequency judgment instead of a probability judgment? Hertwig and Gigerenzer (1999) gave participants a task describing 200 women who fit the description of Linda. They were then asked "How many of the 200 women are bank tellers?", "How many of the 200 women are feminists?", and "How many of the 200 women are bank tellers and are active in the feminist movement?" When questioned this way, no participant violated the conjunction rule. The key here is asking people for a frequency estimate rather than asking them to rank order probabilities. Following up on this, Sloman and colleagues (2003) found almost 70% of participants violated the conjunction rule when asked to rank order-estimated frequencies, but only a little more than 30% made this error when asked to simply estimate frequency in the way Hertwig and Gigerenzer did.

How the Question Is Framed Determines Whether You Get It Right or Wrong

The Linda and Tom problem results show how easily people's judgments can be influenced by changes in wording. More broadly, these results show the impact of the framing effect: People's decisions are influenced more by how the problem is framed (described) than by the objective data contained in the problem. The strongest framing effects are usually found when probability information is pitted against a deep-seated bias in our cognitive architecture.

The best example of this is loss-aversion bias, people's tendency to strongly prefer avoiding losses to acquiring gains (Tversky & Kahneman, 1981). Some studies suggest that losses are psychologically twice as powerful as gains.

Here is an example that brings this point home clearly:

> *Imagine you are a patient with lung cancer. Which of the following two options would you prefer?*
>
> *A. Surgery: Of 100 people undergoing surgery, 90 live through the post-operative period, 68 are alive at the end of the first year, and 34 are alive at the end of five years.*
>
> *B. Radiation Therapy: Of 100 people undergoing radiation therapy, all live through the treatment, 77 are alive at the end of one year, and 22 are alive at the end of five years.*

Forty-four percent of respondents to this scenario favored radiation therapy over surgery. Now read this scenario:

> *Imagine you are a patient with lung cancer. Which of the following two options would you prefer?*
>
> *A. Surgery: Of 100 people undergoing surgery, 10 die during surgery or the post-operative period, 32 die by the end of the first year, and 66 die by the end of five years.*
>
> *B. Radiation Therapy: Of 100 people undergoing radiation therapy, none die during treatment, 23 die by the end of one year, and 78 die by the end of five years.*

Using this scenario, only 18% favored radiation therapy over surgery. There was no difference between patients or physicians in this regard. But take a closer look at the numbers. They are exactly the same: "90 live" is the same as "10 die," "68 are alive" is the same as "32 die," and so on. It's exactly the same problem twice – except that one couches the decision in terms of survival, whereas the other couches the decision in terms of mortality.

This kind of "loss aversion" also strongly influences what people do with their money. Consider this situation:

You have been given $1,000. You now have to choose to either

Take a Risk	*Play it Safe*
Heads = You get $1,000 more.	*You get $500 more.*
Tails = You get $0 more.	

Most people choose to play it safe – take the extra $500 and run! Now take a look at this scenario:

You have been given $2,000. You now have to choose to either

Take a Risk	*Play it Safe*
Heads = You lose $1,000.	*You lose $500.*
Tails = You lose $0.	

Most people choose to bet because they want to avoid losing the $500. But here's the problem: It is exactly the same bet both times. Each offers a fifty-fifty chance of $1,000 versus $2,000 or simply taking $1,500.

To put it more plainly, when faced with possible losses, people choose the risky alternative. In the above example, most people would rather take a bet that could end up costing them double the sure loss. They take the bet because they hope the coin will land tails up. If this were an investment, this means you will hold onto an investment as it loses money, hoping it will regain its value. More likely than not, you will ride it all the way down, as many of us did with our investments and homes during the 2008 economic meltdown.

We are not the only ones who act this way. According to researcher Laurie Santos, Capuchin monkeys do the same thing. She and her co-investigators devised a series of studies in which Capuchin monkeys had to choose between an offering of grapes from two different researchers (Lakshminarayanan and colleagues, 2011). In the gain condition, one researcher always added one grape before giving up the grapes. If his offer was accepted, the monkey always received two grapes. The other researcher sometimes added nothing and sometimes added two more. So if his offer was accepted, the monkey sometimes got three grapes and sometimes got only one grape. In the loss condition, the researchers offered the monkey three grapes. The first researcher always took away one grape before giving them the grapes, so if his offer was accepted, the monkey always ended up getting two grapes. The other researcher sometimes took nothing away, or sometimes took two grapes away. So if his offer was accepted, the monkey sometimes got three grapes and sometimes got only one grape.

Surprisingly, the monkeys in this study acted just like humans: They went for the sure-bet guy in the gain condition, and for the risky guy in the loss condition. As Santos points out, this means that this strategy has been around for at least 35 million years because that is how long ago humans and Capuchins shared a common ancestor on our evolutionary tree.

Kahneman and Tversky (1979) demonstrated the mathematical implications of these human biases in Prospect Theory, a brilliant treatise that won them the Nobel Prize in economics. There are two key mechanisms in this theory: (1) People respond more strongly to losses than comparable gains (winning $10 feels good, but losing $10 feels much worse), and (2) people respond to changes in relative gains and losses rather than absolute gains or losses (losing or gaining $10 means more if you have only $20 than if you have $200). This means that people make decisions by comparing them to a flexible reference point.

Kahneman (2003) also proposed a dual-process theory to explain some of these characteristics of human decision making. The most important point is that dual-process theories distinguish between a rapid decision-making system (system 1) and a slower one (system 2). According to Kahneman, the rapid system outputs decisions based on emotion, intuition, and heuristics (such as causal interpretations of events and prototypes). The slower system 2 outputs decisions based on the abstract nature of the problem, particularly its statistical or logical structure.

When system 1 is in charge, we can end up being overly confident in our decisions. This is referred to as *overconfidence bias*: People (novices and experts) are more confident about their decisions than is justified given the environment in which they are making their decisions. As a result, they frequently stop their search for answers before all available evidence can be collected. Here are some examples:

> *In each of the following pairs, which city has more inhabitants? Also indicate your level of confidence in your decision on a scale of 1 to 10, with 10 meaning perfect confidence and 0 meaning guessing.*

(a) Las Vegas	*(b) Miami*
(a) Sydney	*(b) Melbourne*
(a) Berlin	*(b) Heidelberg*

> *In each of the following pairs, which historical event happened first? Also indicate your level of confidence in your decision on a scale of 1 to 10, with 10 meaning perfect confidence and 0 meaning guessing.*

(a) Signing of the Magna Carta	*(b) Birth of Mohammed*
(a) Death of Napoleon	*(b) Louisiana Purchase*
(a) Lincoln's assassination	*(b) Birth of Queen Victoria*

Table 3.3 presents the results in terms of the confidence and accuracy.

As is apparent, people's confidence ratings were much higher than was warranted by their accuracy. Warning participants that people are often overconfident has no effect, nor does offering them money as a reward for accuracy. This phenomenon has been demonstrated in a wide variety of subject populations including undergraduates, graduate students, physicians and even CIA analysts. (For a survey of the literature see Lichtenstein, Fischoff & Phillips, 1982.)

But before you throw in the towel, consider that demonstrating this kind of overconfidence depends largely on the questions you choose to ask. Juslin, Winman, and Olsson (2000) analyzed 135 studies on the overconfidence effect, and found that items had been chosen in ways that inadvertently overrepresented "trick items" that were likely to lead to the wrong answer. Take the first set of questions concerning

Table 3.3. *The Relationship between Performance Confidence and Performance Accuracy*

Confidence (%)	Accuracy (%)
100	80
90	70
80	60

population size of cities. A fair way to test general knowledge of this domain and confidence in answers given might be to take all German cities with more than 100,000 inhabitants, and then generate a test sample by randomly selecting cities from that pool. When items are chosen this way, overconfidence largely disappears.

Why do we falsely appear to be overconfident in our decisions? Perhaps because the heuristics we use are adaptive in our normal environments, meaning they usually give us pretty good answers. So, in a natural environment, we tend to make accurate decisions. Gerd Gigerenzer's *Adaptive Behavior & Cognition Program* has identified a number of fast and frugal heuristics (based on limited search) that perform about as well as algorithms that require much more information and (in a serial architecture) more time (Gigerenzer et al., 2000). The key is that these heuristics capitalize on environmental regularities to make smart inferences.

The key features of these heuristics are a limited-information search plus a stopping rule. Just as physiological systems have been shaped by a common pressure for *energy*-processing efficiency, psychological information-processing systems have all been shaped by common pressure for *information*-processing efficiency. There is a cost for obtaining and processing information, and there is a benefit for making correct decisions. But in real life, tradeoffs exist between these two. The stopping rule is simply that we stop searching for information when the processing costs equal or exceed the benefits accruing from the additional search. In common parlance, you want to avoid paralysis by analysis, and just get on with making a decision.

In chapter 5 of their 2000 book, *Simple Heuristics That Make Us Smart*, Gigerenzer, Todd, and the Adaptive Behavior and Cognition group describe the results of a study in which different decision models were used to predict which of two objects randomly chosen from the following twenty data sets would score higher on some criterion:

High school dropout rates
Homelessness rates
Mortality
City populations
House prices
Land rent
Professors' salaries
Male attractiveness
Female attractiveness
Car accidents
Fuel consumption
Fish fertility
Mammals' sleep
Cow-manure oxygen
Biodiversity
Rainfall
Oxidants in Los Angeles
Ozone in San Francisco
Car accidents
Fuel consumption
Obesity at age 18
Body fat

Examples of the criteria to be compared are

Which city has the larger population?
Which high school has the larger dropout rate?
Which highway has the larger number of accidents?
Which individual is more attractive?

The decision models included two normative strategies and two heuristic strategies. The normative models were (1) Multiple regression, a statistical technique that predicts values of one variable from multiple variables, using least-squares methods for parameter estimation, and (2) Dawes Rule, a simplified regression that uses unit weights (+1 or –1) instead of optimal weights. (Essentially, it adds up pieces of positive evidence and subtracts negative evidence.) The heuristics included (3) Take the Best, Forget the Rest, which searches cues in order of their usefulness, stopping when one is found that discriminates between the two choices, and (4) the minimalist heuristic, which looks up cues in random order, and stops when a cue is found that discriminates between two objects. The results were stunning – the heuristics were slightly more accurate than the much more time-consuming and processing-intensive normative models! The prediction-accuracy rate for Take the Best was 71% and for Minimalist 65%. In contrast, the prediction-accuracy rate for multiple regression was 68%, and for Dawes rule, 69%. These results show quite dramatically that heuristic decision making need not be inaccurate or inferior to normative decision making.

Your Brain on Decision Making

Decision neuroscience is a discipline that investigates brain activity during decision making (Sanfey 2007). The results of this exciting new field have clarified our understanding of human decision making and provided surprising support for some decision-making theories.

As we saw, rational-choice theory – the granddaddy of decision theories – described rational decision making as the product of probability estimation and utility. It turns out that these two functions are neurologically real and neurologically separable. The part of you that calculates the probability of a choice is separate from the part that calculates how happy that choice will make you.

Let's start with utility. Dopamine is a neurotransmitter (chemical) that is released when a reward is received or in anticipation of getting a reward. Even cues that are associated with a reward will trigger a release of dopamine. Neurons that release dopamine are clustered in four areas in the brain: the nucleus accumbens, the ventral tegmental areas (VTA), the striatum, and the frontal cortex. These areas can be thought of as your brain's reward circuitry. Any activity we find pleasurable (from hearing our favorite music to seeing a beautiful face) activates this system. These circuits enable our brains to encode and remember the circumstances that led to the pleasure so we can repeat the behavior and go back to the reward in the future. When this circuitry is active, it is interpreted as a neural signature of reward (utility) processing. When you receive a reward or when you make a decision that you believe will bring you a reward, these same areas are activated. So, as far as your brain is concerned, experienced utility and "decision utility" feel the same (Breiter, Aharon, Kahneman, Dale, & Shizgal, 2001; O'Doherty, Deichmann, Critchley, & Dolan, 2002).

Box 3.2. Hijacking Your Brain's Reward System

All addictive drugs stimulate the brain's dopamine reward centers far more than usually occurs in everyday life. As a person continues to abuse drugs, the brain adapts to the overwhelming surges in dopamine by producing less dopamine or by reducing the number of dopamine receptors in the reward circuit. As a result, dopamine's impact on the reward circuit is lessened, reducing the abuser's ability to enjoy the drugs and the things that previously brought pleasure. This compels addicts to increase the dose in order to attempt to bring their dopamine function back to normal or to achieve the same "high" (http://www.nida.nih.gov/scienceofaddiction/).

Consider, for example, cocaine. Cocaine blocks the chemicals that normally remove dopamine from synapses after the neuron has been activated. If dopamine lingers in the synapses for longer than normal, it prolongs the stimulation of receptors and causes pleasurable effects. In time, this overstimulation damages or destroys dopamine receptors, reducing their numbers. Soon increased amounts of a drug are required to stimulate the same amount of activity.

As Wilson and Kuhn (2005) put it,

So addiction is far more than seeking pleasure by choice. Nor is it just the willingness to avoid withdrawal symptoms. It is a hijacking of the brain circuitry that controls behavior so that the addict's behavior is fully directed to drug seeking and use. With repeated drug use, the reward system of the brain becomes subservient to the need for the drug.

Knutson and colleagues conducted an fMRI experiment in which both reward magnitude and probability were manipulated (Knutson, Taylor, Kaufman, Peterson, & Glover, 2005). Participants were presented with cues indicating both the likelihood and value of upcoming monetary rewards. They found that activity in the medial prefrontal cortex was related to the subjective *probability* of obtaining the reward, but activation in midbrain areas correlated with expected reward magnitude. Moreover, people's verbal reports of their probability estimates correlated with prefrontal brain activity, whereas their reports of arousal correlated with midbrain activity. Recent studies using fMRI (Breiter et al., 2001) and electroencephalogram (EEG) (Holroyd et al., 2004) also demonstrated that neural signatures of reward are determined by the value of the outcome relative to the range of possible outcomes rather than by the objective value of the outcome itself, as predicted by Prospect Theory.

Results like these show conclusively that the probability-estimation process is neurologically separable from reward assessment. The study also showed that utility-theory models may be an accurate representation of how the brain decides between alternatives. The crucial difference, though, is that standard economic models based on rational choice presuppose the operation of a single rational information processor. Results of neuroscience research indicate that decisions are the outcome of two separate neural processors.

These separate processors also have been found to compete with each other in decision making. When people decide to make risky decisions, the reward areas of the brain become highly active just prior to making the decision. In other words, this neural signature shows that they are anticipating large payoffs and are not thinking about the *probability* of payoffs (Knuston & Bossaerts, 2007). This is one reason gambling can be so addictive; the act of placing the bet can feel as rewarding as winning.

The neural signatures underlying framing effects have also been identified. DeMartino and colleagues (2006) had people make decisions about bets framed as sure or risky gains or losses (as discussed above). They found that when people fall prey to such framing effects, the emotion-processing areas of the brain (e.g., amygdala) are very active. When people do not fall prey to framing effects, these areas show reduced activity, and the anterior cingulate cortex (which signals a conflict between the outputs of the emotion-based and deliberation-based systems) is very active. The researchers pointed out that these results suggest a competition between two neural systems that is detected by ACC activation, an emotion-based system (amygdala) and a predominantly analytic system (orbital and prefrontal cortex). Moreover, people who showed more activation in the orbital and prefrontal cortex were also least affected by frame manipulation.

Finally, De Neys and colleagues (2008) had people make decisions about engineers-versus-lawyers–type problems that pit prototype thinking against probability-based decision making and neutral problems that did not pit the two against each other. Their results showed that the ACC was indeed more active on the conflict-based problems than on the neutral problems. When people overcame their bias to make decisions on the basis of similarity to prototype, the right lateral prefrontal cortex was very active. This means that biased decision making appears to be due to a failure to override intuitive heuristics. In the engineers-versus-lawyers–type problems, system 1 makes a decision quickly by comparing a description to your prototypes for engineers and lawyers. Whichever matches best is the right answer as far as it is concerned. System 2 reaches a decision more slowly by focusing on the problem structure, particularly the probability information given. Its decision will be based on your knowledge of probabilities. When people's final decisions were heuristic-based (based on similarity to prototype), system 1's quick response won the competition. When people made a reason-based decision, system 2 over-rode system 1's quick response.

Decision Making in the Real World: The Economic Meltdown of 2008

It is not a long stretch to say that "rational" self-interest constituted the core of the global economic meltdown that happened in 2008. This is what happened (according to the 2011 report of the Financial Crisis Inquiry Commission): Banks offered easy credit to mortgage-loan applicants because it was profitable for them to do so. They bundled these mortgages and sold them as investments, thereby removing any risk or responsibility for themselves if it turned out that people could not repay them. Real-estate brokers made a cut on each home they sold, so it was definitely in their self-interest to sell homes whether or not the buyers could actually afford them. After all, the mortgage was the bank's problem. Given the availability of easy credit, millions of people took out loans larger than they could afford in the hopes that they could either flip the house for profit or refinance later at a lower rate. And the price of homes grew exponentially because so many people wanted to buy, causing bidding wars.

Everything was great, until it wasn't. People who were given mortgages they couldn't really afford couldn't repay them, particularly those who had taken out adjustable-rate mortgages. These mortgages start out at a low rate and then increase substantially in a few years. People who had been paying $500 monthly to live in their homes suddenly found themselves having to pay $1,500 monthly. And they couldn't afford that. So they put their home on the market along with the millions of others homes that people could no longer afford.

With so many houses on the market at the same time and so many people looking for financing or re-financing, the housing market went bust. People began walking away from their mortgages, leaving their homes for the banks to auction to people who could not get credit because there was no credit to be found. And that set off a chain reaction in the economy.

Many banks and investment firms began bleeding money due to massive losses in mortgage-based investment. The banking behemoth Lehman Brothers went into bankruptcy on September 15, 2008, and others seemed sure to follow. These mortgage-backed securities had been marketed around the world, so the financial crisis was not limited to the United States. This ultimately led to a recession that was the worst since the Great Depression of the 1930s – a financial meltdown also instigated by bad investment decision making. The federal government became worried about a meltdown in the value of U.S.-backed investments and deemed these financial institutions "too big to fail" because they would take down the rest of the world. So it orchestrated the largest federal-based bailout in history. The tax dollars of average citizens were used to shore up the sinking ship of the U.S.-investment-banking industry.

Myopic self-interest may have been the core of this economic meltdown, but it was helped along by the fact that the agents driving our markets are human beings, and, as we've seen, we have a psychology of decision making that combines rather disastrously with market forces. We are frequently required to make decisions based on calculations that are difficult to do in our heads. Throw into the mix the fact that we are frequently deciding under conditions of uncertainty – that is, we frequently have insufficient information either because that's all there is or because information is asymmetrical – the other interested party knows more about an investment than we do. (As P.J O'Rourke pointed out in his book *Eat the Rich*, it is difficult to distinguish between asymmetrical information and insider trading because they are exactly the same thing. It's how Martha Stewart ended up going to prison.) We try to avoid risk, yet we make extremely risky decisions when facing potential losses. If we buy an investment that promptly gains in value, we will frequently sell it to access the gains. But if we buy an investment that promptly loses value, we will frequently hang onto it indefinitely, hoping it will regain its original value. We will ride that bad investment all the way down. We are also prone to overconfidence when it comes to making decisions.

Given all of this, there has been a movement in economics to take into consideration human psychology when describing markets. This field is called behavioral economics, and it carries with it a hope of improving both our predictions about human decision making as well as our ability to prevent its negative consequences.

FOUR

Moral Decision Making

HOW WE TELL RIGHT FROM WRONG

In 1978, philosopher Phillipa Foot asked readers to consider the following moral dilemma:

> *A trolley is running out of control down a track. In its path are five people who have been tied to the track by a mad philosopher. Fortunately, you could flip a switch, which will lead the trolley down a different track to safety. Unfortunately, there is a single person tied to that track. Should you flip the switch or do nothing?*

Most people choose to flip the switch because choosing to save five lives over one life seems like the right thing to do. But now consider this version of the trolley problem proposed by Judith Jarvis Thomson (1985).

> *As before, a trolley is hurtling down a track towards five people. You are on a bridge under which it will pass, and you can stop it by dropping a heavy weight in front of it. As it happens, there is a very fat man next to you – your only way to stop the trolley is to push him over the bridge and onto the track, killing him to save five. Should you proceed?*

In contrast to the standard-trolley problem, most people believe that pushing the fat man is wrong – even though it means five other lives will be lost.

As these thought problems show, our intuitions in moral matters sometimes appear contradictory. To make matters worse, moral issues typically elicit very strong emotional reactions from people and can have wide-reaching impact on the lives of others. The Civil War was fought in large part because the country was divided on the moral issue of slavery, the civil-rights movement gained unstoppable momentum when people began to believe that sociopolitical inequities such as Jim Crowe laws were morally wrong, and the disparate moral reactions to *Roe v. Wade* continue to reverberate through the political landscape, oftentimes determining election results. In light of this strong impact moral judgments have on our lives and the often thorny nature of moral dilemmas, we frequently find ourselves searching for guidance in these matters. This chapter makes no pretense to providing a comprehensive treatment of these weighty matters. The goal here instead is to sketch the most influential treatments offered by secular moral and ethical theorists in Western culture.

Church and State Weigh in on Morality

The first thing to notice is that rational choice theory is not particularly useful here. As we saw, the basic tenet of rational choice is that rational agents always act in their own self-interest. Your self-interest does not appear to be on the line in the trolley problems. The dilemmas involve strangers, your life is not threatened, and there appears to be no direct benefit to either choice as far as you are concerned.

Instead, these problems rely on the implicit concept of a *moral imperative*, an action that must be taken because it is the right thing to do. While writing this chapter, I Googled "moral imperative" and got over a million hits, most of which were headlines like these:

The Moral Imperative of School Leadership
Obama Quotes Scripture to Push "Moral Imperative" on Immigration
The Moral Imperative of Literacy
Shocking Study Results Reveal Moral Imperative to Fix Medicaid
How Killing Libyans Became a Moral Imperative
The Moral Imperative to End Poverty

In each of these cases, the point is that a suggested course of action is not just recommended or even commendable. It is obligatory. Improving literacy isn't just a good choice when cast this way; it becomes a moral obligation. Not improving literacy is judged immoral.

Moral dilemmas like the trolley problems have a special wrinkle to them. *We find ourselves in a moral dilemma when the choices we are facing pit moral imperatives against each other; obeying one would result in transgressing the other.* No matter which course of action you choose to take, you will end up breaching a moral principle. In the standard and fat-man trolley problems, you must decide whether or not to sacrifice one life to save others. This pits two moral imperatives against each other. The first is that it is right to act in a way that save lives. The second is that it is wrong to take a life. Yet our opposite reactions to these dilemmas suggest that there are more moral imperatives lurking about in these problems than are apparent upon first reading. So we want to know what these moral imperatives are and where they come from.

The answers to these questions depend in large part on whether you take a religious or secular view. Religions typically attribute moral imperatives to a divine authority. A *theocracy* is a form of government in which the governing authorities are believed to be divinely guided, and so the laws imposed by these authorities have moral force. This is sometimes referred to as "revealed religion" – that is, the revealing to humans by God ideas that cannot be arrived at through reason. Our founding fathers were profoundly skeptical of forming a government on such a foundation. In a letter to the Danbury Baptists in 1802, Thomas Jefferson put it this way:

> Believing with you that religion is a matter which lies solely between Man & his God, that he owes account to none other for his faith or his worship, that the legitimate powers of government reach actions only, & not opinions, I contemplate with sovereign reverence that act of the whole American people which declared that their legislature should "make no law respecting an establishment of religion, or prohibiting the free exercise thereof," thus building a wall of separation between Church & State.

Along with John Adams and James Madison, Jefferson was much influenced by eighteenth-century Enlightenment philosophers who stressed reason and scientific observation as a means of discovering truth. The most influential of these was David Hume. His ideas were challenged by three very influential nineteenth-century philosophers, Immanuel Kant, John Stuart Mill, and Jeremy Bentham. Together, the writings of these eminent thinkers form the foundation of the modern field of moral philosophy, or ethics.

The questions addressed by *moral philosophy* are those concerning issues of good and evil, right and wrong, virtue and vice, and justice. Moral theory is not concerned with how we *actually* behave, but rather with how we *ought* to behave. Three core concepts underlie all moral philosophy. Permissions are actions we may perform if we wish to (e.g., donate to charity). Prohibitions are actions that are wrong to perform (e.g., kill an innocent person). Prescriptions are actions that are wrong *not* to perform if it is possible for us to perform them (e.g., save an innocent person). These are also cornerstones of legal theory.

What David Hume Had to Say

David Hume (1711–1776) was a Scottish philosopher, economist, and historian. He was a founder of a philosophical school of thought called British empiricism that rejected the possibility of certainty in knowledge. To an empiricist, all knowledge is acquired through experience. Hume laid out his very influential moral theory in book three of *A Treatise of Human Nature* (1740) and in *An Enquiry Concerning the Principles of Morals* (1751). In the first part of book three, he asked the following question (which continues to be a core subject of scientific investigation and philosophical treatises today): Are moral judgments rational judgments about conceptual relations and facts, or are they emotional responses? Hume believed that they are emotional responses. To demonstrate his point, he put forth his famous argument from arboreal patricide: A young tree that overgrows and kills its parent exhibits the same alleged relations as a human child killing his or her parent. So if morality is merely a question of relations, then the young tree is immoral. As Hume pointed out, this is plainly absurd. The crucial point of Hume's analysis is this: We cannot deduce statements of obligation from statements of fact. And since moral approval is not a judgment of reason, it must be an emotional response.

Hume developed his full moral theory around a chain of events in which an agent action impacts a recipient and is observed by a spectator. Hume believed an agent's moral actions are motivated by character traits that could be either virtuous or vicious. For example, people who voluntarily donate money to charity are motivated by a virtuous character trait. Those who steal money from the charity to enrich themselves are motivated by a vicious character trait.

Hume believed that some character traits are natural and others are acquired (or, to use his term, artificial). Using our donation example again, donating to a worthy cause usually makes us feel good inside. So this kind of moral act is grounded in a natural feeling. The recipient may also feel gratitude for the donation, which is also grounded in a positive feeling. Finally, an observer may sympathetically feel the positive emotion of the recipient when observing this act of kindness. If instead, you stole money from the charity, you would feel bad, as would the people you stole from, and any observers who witnessed the theft. *To Hume, these sympathetic feelings constituted a moral judgment of the act.*

Hume also introduced the notion of utility into moral theory; we approve of moral acts in part because they have utility – they are useful. But, in a section titled *Why Utility Pleases* (Section V of Hume, 1751), he argues that we approve of such useful actions because of our ability to sympathize with the recipients of those actions.

By grounding moral judgment in emotion, Hume's theory readily explains why people give different answers to the standard and fat-man versions of the trolley problems: The up-close-and-personal description of the fat-man version elicits stronger emotions than the standard-trolley version's remote-switch description. It is one thing to flip a switch, thereby yielding an unfortunate but beneficial outcome; it is quite another to grab a human being against his will and throw him to his fate. Even though the same number of people are sacrificed and saved in both versions, the emotional impact of these problems could not be more different.

Hume also addressed the more abstract and political concept of justice. He believed the concept of justice is not natural but instead emerges from human convention and is passed on through education.

Because we depend on society to survive, we want to advance society. This means acknowledging our responsibilities toward others who allow us to achieve that end. The three main rules of justice that naturally emerge from these considerations include honoring the stability of possessions, transference of possessions by consent, and performances of promises (contracts). Governments emerge in order to protect us in the agreements we enter into by enforcing them and to protect society as a whole by forcing individuals to make some agreements for the common good.

The highest merit that a human can achieve according to Hume was benevolence. In closing this section, I have chosen to quote here from the section *How Benevolence Is Valued* from book three of *Treatise* (Hume, 1740). The ideas expressed here give you the full flavor of how this giant of the Enlightenment thought about justice and human nature. Notice that he appeals not just to fairness, but to "noblesse oblige":

> You may well think that there is no need to show that the benevolent or softer affections are estimable, and always attract the approval and goodwill of mankind. All languages have equivalents of the words "sociable," "good-natured," "humane," "merciful," "grateful," "friendly," "generous" and "beneficent," and such words always express the highest merit that human nature can attain. When these amiable qualities are accompanied by noble birth and power and distinguished abilities, and display themselves in the good government or useful instruction of mankind, they seem even to raise the possessors of them above the rank of human nature, making them somewhat approach the status of divine. Great ability, undaunted courage, tremendous success – these may expose a hero or politician to the public's envy and ill-will; but as soon as "humane" and "beneficent" are added to the praises – when instances are displayed of mercy, gentleness, or friendship – envy itself is silent, or joins in with the general voice of approval and applause.
>
> When Pericles, the great Athenian statesman and general, was on his deathbed, his surrounding friends – thinking he was unconscious – began to express their sorrow by listing their dying patron's great qualities and successes, his conquests and victories, his unusually long time in power, and his nine trophies erected "to celebrate victories" over the enemies of the

> republic. In fact the dying hero was conscious, heard all of this, and joined in: "You are forgetting the highest of my praises. While dwelling on those common advantages, in which luck had a principal share, you haven't observed that no citizen ever wore mourning because of me."

What Immanuel Kant Had to Say

Kant was a German philosopher who rejected the empiricist view that all knowledge derives from experience. Instead, he argued that reason was the source of all knowledge and justification. This was true, he argued, for morality as well. In stark contrast to Hume's belief in the emotional basis of morality, Kant proposed a theory of morality in which *a moral judgment is the outcome of rational thought.* Kant's moral theory is called *deontology* – a theory of morality that is grounded in duties (rights and obligations). His moral theory can be found in *The Foundations of the Metaphysics of Morals* (Kant, 1785) and *The Critique of Practical Reason* (Kant, 1787). The core concept of his position was the *categorical imperative* – moral rules that are discoverable entirely through reason alone and are absolutely binding for all rational agents.

Like Hume, Kant believed than the moral worth of an action depended on the motivations behind it. But whereas Hume grounded motivation in virtuous or vicious character traits, Kant grounded them in universal principles that are discovered through reason. The concepts of *autonomy* and *universality* are critical to his moral theory.

To understand why autonomy plays such a large role in his theory, you have to understand how Kant conceived of humans in the natural world. He believed that the behavior of animals was entirely causally determined by forces acting upon the animals. They ate, bred, killed, cared for their young, and so on because this behavior was instinctive and triggered by physical causes. For this reason, the concept of morality did not apply to them. A dog is not committing an evil act when it kills a cat, even if the action evokes strong emotions in us. Animals cannot reason or choose how to act, and so they cannot be held morally accountable.

Not so for human beings. We are capable of rational thought, so we can choose how we act, and thus we can be held morally accountable.

Because we can reason, we are *autonomous* beings. Because we can reason, we can choose how to act; our behavior is not causally determined – it is not instinctive or reflexive. The world therefore divides neatly, in Kant's view, into autonomous beings who are ends in themselves (self-governing) and non-autonomous beings (behaving reflexively or instinctively). Our ability to choose is our *free will.*

But here's the catch: If we were purely rational, then we would never make mistakes. But we are neither purely animal nor purely rational. We are somewhere in between, and so can choose wrongly. We sometimes give in to our impulses and sometimes act according to principle. We frequently have a number of behavioral choices at our disposal that constitute better or worse means for achieving our goals. If we need money, one choice is to take it from someone else by force. Another is to borrow the money. A third is to earn it by providing a good or service. *We need rules to tell us how we should choose when we have the power to choose.* We need rules of conduct that tell us how we *ought* to behave.

How do we discover these rules of conduct? Kant believed evaluating rules (or actions) based on their outcomes is a non-starter because we can't control outcomes. Even the best choices can yield unforeseen disastrous consequences. What we can control, however, is our intentions – our motives – underlying the actions. So the morality of an action is a function of the motivations underlying it. To Kant, there is only one motive that can be classified as good without qualification, and that is good will: You intend to do the right thing, and you choose your actions based on that principle. Essentially, it makes no sense to say someone did the wrong thing for the right reason. If the choice was based on right reason, it was the right thing to do. Period.

Similarly, you can't do the right thing for the wrong reasons. Like Hume, Kant would agree that someone who donated to charity because he was required to or because he thought it would bring him better business contacts was not behaving morally. To count as a moral act, the donation had to be freely given because it was the right thing to do. Unlike Hume, however, Kant rejected as moral donation made out of desire to feel generous; he also rejected as immoral not making a donation because you prefer to be selfish. To Kant, these emotions were entirely subjective and therefore irrelevant. What matters is that you act according to right principles, and you choose to do so out of conscious deliberation. If you act this way, you are doing your duty.

So how do we come to know these "right principles," these duties? Kant believed that we find these through pure reason. The first quality of a moral principle is universality: Morality must be the same for everyone – you can't make an exception for yourself or anyone else. *Thus, the test for morality (the test for duty) is whether it can be willed that everyone act in the same way.* Kant believed that reason would ensure universality because the discoveries of reason would be the same for every rational agent.

Let me make plain what he said: Because humans are autonomous beings capable of reason, and because reason will always lead to the same discoveries in all rational agents, then we can discover for ourselves what we ought to do on every occasion for moral choice. We don't need the state, or a religion, or any authority to tell us what to do. We choose for ourselves. And if we make our choice according to right principles arrived at through pure reason, then the act is moral – regardless of its consequences.

Kant distinguished between two kinds of laws that reason can create, hypothetical and categorical imperatives. A *hypothetical imperative* is a conditional rule that expresses an action that may be taken to achieve some end, such as "If you want people to trust you, don't make promises you don't intend to keep." But this won't work as a categorical imperative – a rule that applies to everyone everywhere – and morality must be universally binding. For one thing, it says "if you want people to trust you," but that doesn't apply to people who don't care whether anyone trusts them. Still, it seems that it is immoral for them to make promises they don't keep. It also makes reference to consequences (people trusting you or not), and consequences are irrelevant to Kant. A categorical imperative does not admit of exceptions, subjective desires, or consequences.

To Kant, the way reasoning discovers a categorical imperative is by detecting *inconsistencies – contradictions.* For example, if you think about it, you will immediately realize that something cannot be both a circle and a square at the same time. That would be a contradiction.

"You ought to square the circle" would be a non-starter as a categorical imperative. When you reason about a duty, you ask yourself whether you should will it to be a duty for everyone everywhere. If you can't find any contradictions to that question, you have found a categorical imperative. For example, this wouldn't work as a categorical imperative: "You have a duty to make promises you don't intend to keep." Why? Because if you think about it, you realize that if everyone made false promises, promises would be meaningless; the term "promise" would be a vacuous concept. You couldn't make a promise because such a thing would not exist. So this doesn't make the cut as a categorical imperative because we can't make it universal without creating a contradiction. But "You have a duty to keep your promises" does make the cut; there is no contradiction to this imperative. It is a duty.

Kant gave at least two formulations of the categorical imperative.

1. Act only according to that maxim by which you can at the same time will that it should become a universal law. (Kant, 1785, *Foundations of the Metaphysics of Morals, Akademie,* p. 422)
2. Act so that you treat humanity, whether in your own person or in that of another, always as an end and never as a means only. (Ibid., p. 429)

Some of our moral duties, according to Kant, include a duty to maintain one's own life, a duty to be beneficent when we can, and a duty to secure one's happiness. The second formulation also suggests a reason people respond differently to the standard and fat-man versions of the trolley problems; the fat-man version requires you to use a human being as a means toward an end, and this violates the categorical imperative to treat people with respect. So perhaps Kant expressed explicitly what we believe implicitly – that this action violates a basic human right.

We saw that Hume believed beneficence to be the highest merit to which humans can aspire. Kant also discussed beneficence:

> A fourth, who is in prosperity, while he sees that others have to contend with great wretchedness and that he could help them, thinks: "What

> concern is it of mine?" … But although it is possible that a universal law of nature might exist in accordance with that maxim, it is impossible to will that such a principle should have the universal validity of a law of nature. For a will which resolved this would contradict itself, inasmuch as many cases might occur in which one would have need of the love and sympathy of others, and in which, by such a law of nature, sprung from his own will, he would deprive himself of all hope of the aid he desires. (Kant, 1785, *Fundamental Principles of the Metaphysic of Morals*, as translated by T.K. Abbott)

So, according to Kant, selfishness does not work as a categorical imperative because it precludes our receiving help when we need it. This also explains why we find the trolley problems troubling. From a purely selfish standpoint, we might ask, "What concern is it of mine whether one or five die?" Benevolence constitutes a duty, and failing to honor this categorical imperative constitutes a contradiction.

Two other deontological principles not considered by Kant merit discussion. The first is the *Doctrine of Doing and Allowing*, which states that it takes more to justify doing a harm than simply allowing a harm. Suppose the fat man were already on the track, and you did nothing to rescue him because his death would save the other five lives. This is a lesser offense because you allowed the harm rather than throw him onto the track. But the same can be said for the trolley; pushing the switch violates this doctrine, and that is one of the sources of the dilemma the decision maker faces according to the deontological view.

The second is the *Doctrine of Double Effect*, which states that it takes more to justify doing harms that were intended than harms that were anticipated but unintended side effects. Think about the standard trolley and fat-man trolley. Tossing the fat man violates this doctrine because you intend to kill him, even if your intent is also to save lives. Killing him is an integral part of your plan to save the others. In standard trolley, your intention in flipping the switch is to save lives, not to kill the lone person on the track. Killing that person is not a necessary part of your plan, but an unintended side effect. So the fat-man scenario violates the Doctrine of Double Effect, the Doctrine of Doing and Allowing, and the second formulation of the categorical imperative. No wonder people say "no" to its moral permissibility!

Finally, it would be amiss to close this section on Kant without discussing the murderer at the door. Kant's focus on categorical imperatives led him to hold certain positions that most of us find, well, nutty. His response to Benjamin Constant's dilemma of the murderer at the door is one such case. Here is the dilemma:

> *Suppose someone asks you to hide him or her from someone else who intends to murder him or her, and you do. Then the would-be murderer comes to the door and demands to know whether the person is there. Do you tell the truth, or do you lie?*

Kant wrote a response to Constant, insisting that even under these circumstances we must obey the categorical imperative that forbids us to tell lies. We must, out of duty born of reason, tell the truth and put this person's life at risk. Remember, to Kant, consequences were irrelevant. As long as we act according to reason-based principles, we have done the right thing. The consequences are out of our hands. Most people reject this rigid application of the duty to tell the truth because we cannot ignore the consequences that our action might well have – ending a life. And the categorical imperative of preserving life (or not killing) trumps the duty to tell the truth.

What Jeremy Bentham and John Stuart Mill Had to Say

Jeremy Bentham (1748–1832) was an English philosopher and jurist who proposed a moral theory known as *utilitarianism* in *Principles of Morals and Legislation* (Bentham, 1789). Mill (1806–1873), an English philosopher and economist, adopted and extended this theory. His most notable works include *On Liberty* (Mill, 1859), which argued for the importance of individuality; *Utilitarianism* (Mill, 1861), which extended Bentham's theory, and *Subjection of Women* (Mill, 1869), which championed women's rights. (In 1866, Mill became the first person in Parliament to call for women to be given the right to vote.) The core principle of utilitarianism is that a right act or policy is the one that causes "the greatest good for the greatest number of people." This "greatest happiness principle" is grounded in the concept of utility.

Although Hume was the first person to introduce the notion of utility into moral theory, Bentham grounded this notion in the experience of pleasure and pain. As he put it in the introduction to *Utilitarianism in Principles of Morals Legislation* (Bentham, 1789),

> Nature has placed mankind under the governance of two sovereign masters, pain and pleasure. It is for them alone to point out what we ought to do, as well as to determine what we shall do. On the one hand the standard of right and wrong, on the other the chain of causes and effects, are fastened to their throne. They govern us in all we do, in all we say, in all we think.

Bentham's principle of utility can therefore be summarized this way: The only good is pleasure, and the only bad is pain. When judging the moral value of actions, we ask whether it causes pleasure or pain to the recipient. What makes a course of action prescribed is that it promotes pleasure; what makes it prohibited is that it causes pain. To be a moral agent means to always act in such a way as to promote pleasure and avoid pain for those whose interests are affected by your action. Mill adopted Bentham's principle as the foundation of his own theory and defined it this way: *Actions are right in proportion as they tend to promote happiness, wrong as they tend to produce the reverse of happiness.* Mill and Bentham's utilitarianism is one type of consequentialism, a branch of philosophy that evaluates the morality of an action on the basis of its consequences.

Both Bentham and Mill were motivated to ground morality empirically in consequences in large part because they opposed ethical intuitionism. Intuitionists hold that there are objective moral truths and that these are recognized by the mind as being immediately self-evident through a faculty of intuition. Bentham and Mill objected to this formulation of morality because, they reasoned, if the validity of moral rules can be intuited, this means that they are *incontestable.* One may therefore simply assert and re-assert moral prejudices indefinitely, and do so without giving any reasons for them. The opportunity for despotism is clear; if I have enough power, my intuitions rule, regardless of their consequences for others. The same goes for Kant's ethical rationalism. Mill argued that Kant failed to show that imperatives were rejected because they led to contradictions. Instead, he pointed out, Kant rejected imperatives when they led to undesirable conclusions.

Going back to our trolley problems, a strict utilitarian would decide to flip the switch and to push the fat man. In both cases, the greatest happiness for the greatest number would be achieved.

What Seems Right to Us: The Psychology of Moral Judgment

When people are given moral dilemmas like the trolley problems, they definitely consider more than just number saved versus number killed. Over dozens of studies, 80% of people say "yes" to flipping the switch and "no" to pushing the man; when given the following problem, they tend to split almost fifty-fifty (Greene and colleagues, 2001):

> *Enemy soldiers have taken over your village. They have orders to kill all remaining civilians. You and some of your townspeople have sought refuge in the cellar of a large house. Your baby begins to cry loudly. You cover his mouth to block the sound. If you remove your hand from his mouth, his crying will summon the attention of the soldiers who will kill you, your child, and the others hiding out in the cellar. To save yourself and the others you must smother your child to death. Should you smother your child in order to save yourself and the other townspeople?*

Here, again, you have to choose whether to sacrifice one life to save many, just as in the trolley problems. But again, people seem to be responding to something more than just number saved or killed. (If you are a *M*A*S*H* fan, this problem may seem familiar; it was the story line for one episode.)

Literally hundreds of experiments have been conducted to investigate people's judgments on dozens of moral dilemmas. Some, like the trolley and crying-baby dilemmas, described saving or sacrificing lives. Others describe less dire circumstances, such as lying on a resume to get a job or cheating on a spouse. We can predict how people will decide based on their decision-making-style preferences as well as their prior moral commitments (Lombrozo, 2009). People who prefer to rely on intuition tend to say "no" to problems like those above, whereas those who prefer to rely on deliberation tend to say "yes" (Hofman &

Baumert, 2010). That is, those who prefer to deliberate when making a decision are more likely to make utilitarian decisions. Not surprisingly, people whose prior philosophical commitments fall into the utilitarian camp tend to say yes, whereas those who fall into the deontological camp tend to say no.

Those results may not surprise you, but these probably will: If we reduce the amount of time that people have to make a decision, they become less utilitarian. For example, the percentage of people who choose to flip the switch drops from 80% to 70%; the percentage of people who choose to smother the crying baby drops from 45% to 13%. So it apparently takes more decision time to approve these actions (Cummins, 2011).

Even more surprising is the impact of seemingly irrelevant factors on moral judgment. Fewer people are willing to flip the switch if they are shown the fat-man problem first. And the reverse is true; more people are willing to push the fat man if they are shown the standard trolley first. This means that people's judgments vary depending on what they were thinking about before making that judgment. This is true not just of average folk; professional philosophers do the same thing (Schwitzgebel & Cushman, in press). (This makes one wonder whether legal judgments vary depending on the types of cases heard prior to deciding.) Even more surprising are the impact of factors that seem more irrelevant than judgment order: People are more likely to judge an action as immoral if they are making their decision in a disgusting environment (Schnall and colleagues, 2008a, b), such as sitting next to a used Kleenex!

Just as in other types of decision making, the two leading explanations for these moral-reasoning effects distinguish between two systems of reasoning. The first is psychologist Jonathan Haidt's social intuitionist theory, which (like Hume) puts intuition front and center as the core of moral judgment, with reasoning happening after the fact (Haidt, 2007). The second is neuroscientist Joshua Greene's dual-process model, which describes moral judgment as the outcome of a competition between emotion or intuition and deliberative reasoning (Greene, 2007).

Haidt's dual-process theory is entirely consistent with Hume's view of moral judgment. Moral intuition is the domain of system 1; it delivers judgments that are fast, automatic, and usually emotion-driven. The nature of the judgment is a simple evaluative feeling of good-bad or like-dislike about the actions or character of a person. These types of judgments appear in consciousness without any awareness of having gone through steps of search, weighing evidence, or inferring a conclusion. That instant reaction *is* our moral judgment.

Deliberative reasoning is the domain of system 2. According to Haidt, it is invoked after the fact, usually to justify a judgment already made. This is a controlled, less emotional process that depends on conscious mental activity. It involves processing information about people and their actions in order to reach a moral judgment or decision. So to Haidt, moral judgments are like aesthetic judgments: We see an action or hear a story and have an instant feeling of approval or disapproval. We can't explain why, but we can cook up a reasonable justification for judgments we've already made.

As evidence, Haidt cites the phenomenon of moral dumbfounding. When people are morally dumbfounded, they know intuitively that an action is wrong, even when they cannot explain why. For example, imagine that a brother and sister slept together once; no one else knew, they used birth control, no harm befell either one, and both felt it brought them closer as siblings. When people are asked whether this was morally permissible, the overwhelming majority say no. But they can't explain why.

Neuroscientist Joshua Greene also appeals to dual processes in explaining moral judgment. But he takes the view championed by Kahneman and others that these processes compete with each other for ascendance in yielding the final decision. If the outputs of the two systems are the same, the final judgment is easy and quick. But if the outcomes differ, the conflict must be resolved. In this case, decisions are slow and difficult, and either process can override the other to yield the final decision. You will feel conflicted or of two minds when making such a decision. Since our first response is emotional, it has precedence. Reason needs time to come on board, and when it does, it will compete with that initial emotional response for control of the final decision. The final decision will simply be whichever system wins.

Greene based his theory on numerous studies in which people made decisions about moral dilemmas while undergoing fMRI imaging of their brains (Greene and colleagues, 2001 and 2004). The first thing that neuroscientists noticed was that, as Hume theorized, the ability to sympathize – or empathize – is a foundation of moral decision making. For example, the brain areas that are active when you experience pain firsthand are the same areas that are active when a normally developing child or adult sees someone else experience pain. Whereas several areas are involved in the experience of pain, the area of most importance to moral empathy is the ventromedial prefrontal cortex (VMPC).

Heekeren and colleagues (2003) had people read statements such as "A steals a car" and "A admires a car," and judge whether or not they were morally permissible. They also had them read non-moral statements such as "A takes a walk" and "A waits a walk," and judge whether they were semantically appropriate. VMPC was far more active during moral as opposed to semantic judgment. Moll and colleagues (2001) found similar results when they had people make silent right or wrong judgments of simple moral statements (e.g., "We break the law when necessary") and non-moral statements (e.g., "Stones are made of water"). Again, activation of VMPC was specific to the processing of the moral statements. When VMPC is activated, it is interpreted to mean that people are responding to the socioemotional aspect of a moral situation.

Greene and colleagues (2001 and 2004) went one step further by imaging people's brains while they reasoned about entire moral dilemmas like the trolley and the crying baby, not just pictures or individual statements. More activation of VMPC occurred when people made decisions about moral dilemmas than when they made decisions about non-moral problems, indicating that people found the moral dilemmas more emotionally arousing. But when people made a utilitarian judgment, brain areas associated with cognitive conflict (anterior cingulate cortex) and abstract reasoning (dorsolateral prefrontal cortex, or DLPFC) were highly active. Greene argued that this pattern of results showed that emotion and reason were neurologically separable during moral decision making; VMPC emotional responses occur first, and making a utilitarian judgment required a different area of the brain (DLPFC) to override that initial emotional or intuitive response.

Koenigs and colleagues (2007) contrasted the moral judgments of patients who had damage to VMPC, patients with damage to other areas of the brain, and non-brain-damaged individuals, all matched for age, gender, and other factors. The VMPC patients were overwhelmingly more likely to give utilitarian responses to a variety of moral dilemmas. One interpretation of these results is that when emotion is kept out of the way, people become more reasonable. Another interpretation is that you have to sustain brain damage to be a strict utilitarian. Either way you look at it, though, it appears that emotion and reason are indeed neurologically separable and can become dissociated through brain damage.

And what about Kant's deontological considerations? Behavioral and neuroscientific findings indicate that these factors weigh in on moral judgments as well (Borg and colleagues, 2006). People judge an action less permissible if the harm is intended than if it is unintended, and they judge taking an action as less morally permissible than doing nothing even if the consequences are the same (e.g., five people die either way). But these are not pure deliberative judgments, contrary to a deontological view. Dilemmas that involve intentional or unintentional harm (violation of the Doctrine of Double Effect) recruit substantial activity in the emotional VMPC (system 1). Dilemmas involving violations of the Doctrine of Doing and Allowing recruit system 2 preferentially only if the consequences of taking the action are the same as not taking the action. If the consequences differ, more system 1 activity (VMPC) takes place. So it appears that Kant was right that moral imperatives underlie moral judgment, but he was also right that we are not purely rational deciders.

Finally, in another study, Young and colleagues (2010) delivered a short burst of magnetic energy to an area of the brain that is associated with our ability to understand what another person is thinking and feeling (right temporoparietal junction). This technique is called transcranial magnetic stimulation (TMS), and it temporarily stuns the area of the brain, rendering it incapable of doing its job. This is like a photoflash temporarily stunning cells on your retina, thereby creating a blind spot in your visual field. In this experiment, TMS disrupted people's ability to take into consideration an agent's intention when judging whether an attempted harm was morally permissible. For example, in one story, a woman named Grace thinks a white substance is sugar and puts it in her friend's cup of tea. In another, Grace thinks the white substance is poison and puts it in her friend's cup of tea. In some versions, the friend dies; in others, the friend is fine. Most people judge the moral permissibility of the action based on what Grace thought the white substance was. If Grace thought it was sugar, then she didn't do a bad thing even if her friend died. If Grace thought it was poison, then she did a bad thing even if her friend was fine. But people who received TMS judged Grace's behavior mostly on whether her friend died or not. Grace's intention held less sway when judging the moral value of her action.

Yes, but What Is Morality for?

Moral issues grab our attention, recruit substantial brain processing, and trigger strong emotional responses. So you would think that they must be serving some very important function. Drawing on the work of French philosopher and sociologist, Emile Durkheim (1858–1917), Haidt points out that morality serves a vital social function: Morality binds and builds; it constrains individuals and ties them to each other to create groups that are emergent entities with new properties. A moral community has a set of shared norms about how members ought to behave, combined with means for imposing costs on violators and/or channeling benefits to cooperators.

This talk of costs and benefits to cooperators should sound familiar; it loomed large in explaining the evolutionary origins of people's choices in games, choices that frequently depart from game theoretical analyses. But Haidt argues that there are other aspects of morality that have equally long evolutionary roots. Based on cross-cultural research, Haidt and Joseph (2008) proposed their *Moral Foundations Theory*, in which they argue that there are five psychological foundations to our moral cognition, each with a separate evolutionary origin. The motivation behind this theory was to explain why morality varies so much across cultures yet still shows so many similarities and recurrent themes. The core of this theory is the proposal that there are five innate and universally available psychological systems are the foundations of "intuitive ethics." Cultures construct virtues, narratives, and institutions on top of these cognitive foundations, thereby creating cultural diversity in moral prescriptions. The foundations are:

1. **Harm/care**, which is rooted in mammalian emotional attachment systems. These systems underlie our ability to empathize with the pain of others. This foundation underlies virtues of kindness, gentleness, and nurturance.
2. **Fairness**, which is founded on reciprocal altruism (discussed in Chapter Two). This foundation generates ideas of justice, rights, and autonomy.
3. **Ingroup/loyalty**, related to our long history as tribal creatures able to form shifting coalitions. This foundation underlies virtues of patriotism and self-sacrifice for the group. It is active any time people feel that it's "one for all, and all for one."
4. **Authority/respect**, which traces it roots to our long primate history of hierarchical social interactions. This foundation underlies virtues of leadership and followership, including deference to legitimate authority and respect for traditions.
5. **Purity/sanctity**, which is related to emotions of disgust that underlie survival-enhancing behaviors such as avoidance of contaminated (bad-tasting) food. This foundation underlies religious notions of striving to live in an elevated, less carnal, more noble way.

According to Haidt, morality everywhere is based on these foundations, but cultures differ according to which ones are emphasized. Interestingly, the moral domain of educated Westerners has been found to be much narrower than it is in other cultures, focusing primarily on harm and fairness. But even more intriguing are results showing that the breadth of moral concerns differ quite substantially depending on political orientation (Haidt & Graham, 2007). Overall, those of a liberal bent focus almost exclusively on harm and fairness when judging the moral merit of an action or policy. This is succinctly summarized as "Does it hurt anyone?" and "Is it fair to everyone?" Conservatives, on the other hand, weigh all five aspects almost equally. To a conservative, violations of respect or purity matter as much as fairness or harm. It is because of these differences in moral outlooks that conservatives and liberals frequently fail to agree on vital matters of social policy. They are essentially speaking from entirely different moral frameworks. As a result, they end up talking past each other. A conservative cannot understand why a liberal does not see that, for example, burning the American flag is wrong. To the conservative, this violates the foundations of respect and in-group loyalty. To a liberal, these foundations are not even on the radar.

FIVE

The Game of Logic

Suppose there is a movie playing at a distant theater that you really want to see. But it is too far to walk there, and you don't have a car. Your deliberation on the matter may go something like this:

I want to go to the movie.
But I don't have a car.
My friend has a car and is going to the movie.
If he takes me along, then my problem is solved.
So I'll call him and ask him to take me along.

The end result of this line of reasoning is an action – calling your friend and asking for a ride. This kind of thinking is called *practical reasoning.* It is reasoning directed toward action.

The thing about the human mind, however, is that it is always thinking – even when we are not trying to solve a problem. More often than not, the end result of this thought process is a belief, not a plan for action. This kind of reasoning is called *theoretical reasoning* (or *discursive* reasoning). *We use theoretical reasoning to determine which beliefs follow logically from other beliefs.* Suppose you asked your friend for a ride to the movie, but he told you his car wasn't working. Then you got a ride from someone else, and you saw your other friend drive into the theater parking lot. You might end up thinking something like this:

He told me his car wasn't working.
But he just drove into the parking lot.

If his car wasn't working, he couldn't drive it here.
So either he fixed it, or he was lying to me.
There wasn't enough time to fix the car.
So he was lying to me.
If he was lying to me, he's not much of a friend.
So he's not much of a friend.

Notice that there is no action-oriented goal to this reasoning. It is instead just a series of inferences concluding in a belief. Sometimes we engage in reasoning deliberately, and sometimes it just happens automatically. In fact, try to stop yourself from weaving logically connected thoughts right now. Don't think about what you're going to do when you're done reading this chapter, or why your friend didn't call you last night. I dare you to try because I know I will win.

Given that you engage in reasoning almost ceaselessly during your waking hours, you should be curious about how it's done. The fundamental unit of thought is a proposition – a statement that may be asserted or denied. Even an image or feeling is, at bottom, a kind of proposition. It has a meaning that can be asserted, denied, combined, and connected logically to other images or feelings. If you imagine Jack and Jill climbing up a hill, the meaning of this image can be expressed propositionally as "Jack and Jill climb up a hill." If I feel particularly fond of you, the meaning of my feeling can be expressed as "I love you."

Propositions are distinct from the sentences that convey them. The sentence "Up a hill climb Jack and Jill" expresses the same proposition as "Jack and Jill climb up a hill." The sentence "You are loved by me" means exactly the same thing as "I love you."

The sentence "I love you" also can be used to express many different meanings (propositions) depending on who the pronouns "you" and "me" represent. It could mean "Denise loves Robert" or "Angelina loves Brad" depending on who is thinking or saying "I love you." But the material point is that in each of these cases, *the proposition expressed is either true or false.*

When we reason something through, we develop a series of propositions that are logically connected to each other. We do the same thing when we try to persuade another person to adopt our beliefs. The material point here is the term *logically connected*. Propositions that are not logically connected to each other may be amusing, but they aren't persuasive. In fact, we may speak of such language disparagingly as "word salad" or "raving". When people stop making sense to us (or we stop making sense to ourselves), we tend to get very worried. In fact, we might wonder whether it's time to call in professional help, such as a psychiatrist.

So what does it mean for propositions to be logically connected? The chief concern of logic is how the truth of some propositions is connected to the truth of others. A set of logically connected propositions is called an *argument*. More precisely, an argument is a set of two or more propositions related to each other in such a way that all but one of them are supposed to provide support for the remaining one. The supporting propositions are premises. The final one that is supported is the conclusion. Here is a simple argument:

PREMISE: *Jack and Jill climbed up the hill.*
CONCLUSION: *Therefore, Jack climbed up the hill.*

The premise in this simple argument is supposed to provide support for its conclusion. This is a pretty good argument. In fact, it is deductively valid – accepting the premises and rejecting the conclusion would constitute a contradiction. If you believe the premises are true, then you are logically committed to the conclusion of a deductively valid argument.

The transition or movement from premises to conclusion – the logical connection between them – is the *inference* upon which the argument relies. You don't need a logician to see the inferential connections between the propositions spawned by seeing your friend arrive at the movie theater in his purportedly dead car. But what about the chain of inferences needed to understand the following story?

Mary put the picnic supplies into the trunk of her car.
The trip took more than an hour.
"Oh, no," she thought ruefully. "The beer will be warm!"

To someone who knows nothing about picnics, this wouldn't be a story at all. It would be three unrelated sentences – veritable word salad. But you probably had no difficulty understanding this story because you do know about picnics. So you filled out this story structure with other propositions that were logically connected, such as:

Mary put the picnic supplies into the trunk of her car.
+*Picnics take place in summer.*
+*The weather is hot in summer.*
+*Picnic supplies include beer.*
The trip took more than an hour.
+*Items stored in trunks in hot weather get warm.*
"Oh, no," she thought ruefully. "The beer will be warm!"

The propositions with plus signs in front of them are propositions that you automatically added to the story based on what you know about picnics. For someone who knows about picnics, this story constituted a set of logically connected propositions. In fact, it constituted a deductively valid argument. To accept all of the premises as true but reject the conclusion (the last line of the story) as false would constitute a contradiction. When an argument is deductively valid, the truth of the premises guarantees the truth of the conclusion.

Notice that this definition depends on the premises being true. Another way to define deductive validity is if the premises are true, then the conclusion must be true as well. Consider this argument:

PREMISE: *The moon is made of Swiss cheese.*
PREMISE: *All Swiss cheese is made at dairy farms.*
CONCLUSION: *The moon is made at dairy farms.*

This argument is deductively valid – that is, if both of the premises were true, then the conclusion would be true as well. To be more precise, we would say that this argument is deductively valid, but it is not sound. A *sound* argument is a deductively valid argument based on true premises.

When an argument claims merely that the truth of its premises make it *likely* or *probable* that its conclusion is also true, it is said to involve an *inductive* inference.

An inductive argument succeeds whenever its premises provide some legitimate evidence or support for the truth of its conclusion. But it would not be completely inconsistent to withhold judgment or even to deny it. For example, analogical reasoning depends on inductive inference. An analogical argument goes like this:

PREMISE 1: *Object X and object Y are similar in having properties Q_1 through Q_n.*
PREMISE 2: *Object X also has property P.*
CONCLUSION: *Object Y also has property P.*

Notice that accepting the truth of the premises does not guarantee that the conclusion must be true as well. Inductive arguments are not evaluated on the basis of validity. Instead, they are evaluated in terms of the strength of the evidence presented in the premises. A strong inductive argument is one whose conclusion is based on a lot of solid evidence. A weak one is based on little or weak evidence.

A Journey into Logic Land

So far, I've been throwing around the term "logically connected," but I haven't really defined it. To do that, we need to understand what logic is.

Logic is a branch of mathematics. In mathematics, we express mathematical relationships using symbols, and we manipulate those symbols according to rules that preserve those mathematical relationships. Here's an example. Suppose you have two apples. Then someone gives you another apple. How many do you have? You could just count them, or you could do some math (actually, arithmetic). You could represent the two apples with a really cool symbol like this one: 2. Then you could represent getting another apple with another really cool symbol, like this one: +. Then you could represent that additional apple with another really cool symbol, like this one: 1. Since you want to know how many apples you've now got, you could represent the equality of the event with another really cool symbol, like this one: =. Finally, you could represent the actual outcome with another really cool symbol, like this one: x: Now, we can express the whole event this way: 2 + 1 = x. We are now in math land where only symbols and rules exist.

To find the answer, we have to make sure that the symbols we chose have the meanings we want. In this case, the symbols "1" and "2" represent whole numbers on the number line. They also represent quantities. Then we need some rules for combining these symbols in ways that get us the right answer. In this case, we need a rule for "+," and the one that works is the addition function. Functions map sets onto one another such that the elements of the sets stand in one-to-one or many-to-one relation to each other. The addition function takes two or more symbols from an input set and maps them onto one and only one symbol in the output set. In this case, it maps "2 + 1" onto the symbol "3".

Are we done? Nope. Now we have to exit math land by translating the output symbol onto a real-world interpretation. We were talking about apples, so we interpret the symbol "3" as referring to the quantity, three apples. Now we're done.

So there are three steps in this game:

1. Translate a real-world event into symbols in math land.
2. Apply the appropriate rule from math land to those symbols in order to get an answer.
3. Take that answer and translate it back into the real world.

At this point, you might be saying, "That's a whole lot of work for just figuring out that two apples plus another apple equals three apples." And you're right. But what if you had 2,378,425 apples, and I gave you 45,823 apples more? Counting to find the answer would be too much work. But if you knew how to translate this problem into math land, it would be very easy – so easy that it could be automated. Any handheld calculator could do it in a few milliseconds.

Here's the leap you need to make: *A logic seeks to do with sentences what mathematics does with numbers – reduce them to symbols that can be manipulated via rules.*

The simplest form discussed here is truth-functional logic, sometimes referred to as zeroth-order logic. Just as mathematical symbols such as + and = stand for the mathematical functions of *addition* and *equality*, in truth-functional logic, symbols such as ~ and ⊕ stand for the logical functions of "negation" and "exclusive or." The arithmetic function + took numerals representing numbers as its input and outputted a numeral that represented the sum of those inputs. In truth-functional logic, the functions take symbols representing propositions as inputs, and they output the truth value of the proposition. A proposition's truth value can take only two values, "true" or "false."

In the bumming-a-ride example discussed above, your thoughts went like this:

Either he fixed his car, or he was lying to me.
There wasn't time to fix the car.
So he was lying to me.

Was this a valid argument? If you confronted your friend with this bit of reasoning, would this constitute a slam-dunk, or could he accuse you of being illogical? We can use truth-functional logic to settle the matter by following the same three steps that we did to find out how many apples we had. In the math problem, we translated the apple quantities into symbols. Here, we translate the propositions into symbols. Just as we had the symbol 2 stand for "two apples," we'll let *P* stand for "suddenly fixed his car" and *Q* stand for "he was lying to me." Then we will use ⊕ to represent "or," ~ to represent "not" (negation), and ∴ to represent "therefore." So now we have represented this argument symbolically in truth-functional logic land:

$$\begin{array}{l} \mathrm{P} \oplus \mathrm{Q} \\ {\sim}\underline{\mathrm{P}} \\ \therefore \mathrm{Q} \end{array}$$

Recall that once we translated the arithmetic word problem into symbols in math land, the numerical symbols 2 and 1 could stand for anything – two apples, two computers, two world wars, and so on. The symbol + stands for one thing and one thing only: the addition function, which outputs the sum of its inputs. Using the addition function is how we solved the problem. In truth-functional logic land, *P* and *Q* now can stand for any simple proposition – "he suddenly fixed his car," "Fido is a dog," "the moon is made of green cheese," "I like chocolate," and so on.

Now we will apply the truth function for "or" (exclusive "or"), which looks like this:

P	Q	P ⊕ Q
T	T	F
T	F	T
F	T	T
F	F	F

This truth function says that a statement of the form is true only when one proposition is true and the other is false. This is what we usually mean by "or" in normal conversation. When the waitress asks you if you want soup or salad, she means you can't have both. You have to choose one or the other.

Now we apply the truth function for negation. This truth function is easy; it just takes a proposition and outputs its opposite. It looks like this:

P	~P
T	F
F	T

Now let's see if this argument is valid using these truth functions:

Propositions		Premises		
P	Q	P ⊕ Q	~P	∴Q
T	T	F	F	T
T	F	T	F	F
F	T	→T	T	T←
F	F	F	T	F

A valid truth-functional argument is simply one where every truth table row that makes its premises true also makes its conclusion true.

So if the statement "P or Q" is true, and the statement "P is false" is also true (third line of the table), then Q must be true. So you win – the argument is valid.

Truth-functional logic has symbols that represent many of the concepts we express in ordinary natural language, such as *and*, *or*, *but*, *if*, *only if*, *unless*, *not*, and *yet*. These functions are used to test validity of arguments in the same way as our exclusive "or" example.

In first-order logic, we can use *quantification* to get inside these propositions in order to express them in terms of predicates, like this: F(j,c). This expression says "John fixed the car" because *F* stands for "fixed," *j* symbolizes "John," and *c* symbolizes "car." This proposition says "fixed(john, car)." We can also use symbols that represent sets of objects or events, such as "Someone fixed the car." This is how we do it:

$$\exists (x)F(x,c)$$

The symbol $\exists$ is called an existential quantifier; it essentially means a set of objects exists that has at least one item in it. So we would read the above expression as "There exists at least one thing (or person) such that it (he or she) fixed the car." We can even express something like "Everyone fixed the car":

$$\forall (x)F(x,c)$$

The symbol $\forall$ is called a universal quantifier and symbolizes "everything," "everyone," or "all." If we wanted to express the idea "All dogs are animals," we would do it like this:

$$\forall (x)(Dx \supset Ax)$$

This says "Take anything you like – if it's a dog, then it's an animal."

Again, the take-home message here is that once you are in logic land, we lose all of the meaning (or content) of the statements. Just as the plus function is "blind" to everything except the shape of the symbols when it maps them onto other symbols, the $\oplus$ function is blind to everything except the shape of the symbols it maps onto other symbols. Once we are done manipulating the symbols using functions, we have to translate these symbols back into their real-world meanings.

Higher order logics have been devised to capture ideas that cannot be expressed in truth-functional or first-order logic. For example, these logics can't express "it is necessary that" or "it is possible that." Yet we say these things all the time and use them in arguments. *Modal logic* allows us to express necessity using the symbol □ and possibility using the symbol ◊. The validity of arguments containing these ideas can't be tested using truth tables because truth tables capture only truth functions. Instead, modal arguments are tested using a possible-world formalism; essentially, you construct a set of possible worlds and then ask whether a particular proposition is true in all or any of them. If it is true in all of them, then the proposition is necessarily true. If it is true in only some, then it is possible. If it is never true, then it is necessarily false. Consider, for example, the proposition "Circles are round." This will be true in any possible world you care to construct. But "circles are red" can be true in some worlds and false in others. So "circles are round" is necessarily true in any argument you care to construct, whereas "circles are red" is only possibly true in any argument you care to construct. "Circles are square" is not true in any possible world, so it is necessarily false in any argument you care to construct. Then you do with the possible worlds what you did with truth tables: An argument is valid just in case any model where the premises are true, the conclusion is true as well.

Finally, there are meanings that require special modal logics. These tend to be domains that are theory-laden. Consider, for example, this argument:

> *If the brake is pressed, then the car will slow down.*
> *The brake pedal is pressed.*
> *Therefore, the car will slow down.*

If we were to express this as a truth-functional argument, it would be valid. This type of argument is called Modus Ponens, and it is always valid due to the truth function for conditionals of the form "if-then."

Propositions		Premises		
P	Q	$P \supset Q$	P	$\therefore Q$
T	T	**→T**	**T**	**T←**
T	F	F	T	F
F	T	T	F	T
F	F	T	F	F

As you can see from the truth table, when both premises are true, the conclusion is true as well. But does this make sense from a causal standpoint? What if the brake lines have been cut? Or the road is very slippery? Or there is no brake fluid? Or the brake shoes are worn? There are many factors that must be taken into account before endorsing this inference. Furthermore, the "if-then" premise in this argument isn't a simple truth-functional conditional. It describes a causal relationship, and causality can't be captured by a truth function. So when we reason causally, we are doing something much more sophisticated that can't be captured adequately with truth-functional logic. We are doing something more like possible worlds, with a special twist.

Causal logic is a branch of modal logic that concerns causal relations (more on this in Chapter Six). Causal relationships are expressed in terms of causal necessity (this factor must be there for the effect to occur) and causal sufficiency (if this factor is present, it is guaranteed that the effect will follow). In this example, pressing the brake is neither necessary nor sufficient to slow a car down. Consider also this argument:

> *If you have a valid library card, then you can take a book out of the library.*
> *You don't have a valid library card.*
> *Therefore, you cannot take a book out of the library.*

In truth-functional logic, this is an invalid argument. (You can trust me on this or work out the truth table to see for yourself.) But it certainly seems like a cogent piece of reasoning. That is because this argument is about permissions and obligations. This is the realm of *deontic logic*, a branch of modal logic that studies the permitted, the obligatory, and the forbidden, which are characterized as deontic modalities (from the Greek *deontos*, "of that which is binding"). Two new symbols are introduced in this logic. *O* stands for obligatory, and *P* stands for permitted. These are interdefined: If you are obligated to do something, you are not permitted to not do it. And if you are permitted to do something, you are not obligated to not do it.

Logicians invent new kinds of logics when the logics available are insufficient to capture legitimate inferences. As the logics become more sophisticated, more complex arguments can be evaluated. For simple arguments like those used as examples here, it may seem like a lot of work for very little benefit. But it seemed the same way when we were considering "two apples plus one apple." The need for symbolic math became vividly clear when we considered 2,378,425 apples plus 45,823. The same is true for argumentation. It can be difficult to keep track of inferences in very long and complex arguments. When you encounter one, translating it into symbolic-logic land and applying the rules of logic can make things immensely easier.

Just How Logical Are People, Really?

So just how good are people at deductive reasoning? Here is a study that shows the typical results of studies aimed at answering that question (Evans and colleagues, 1983). These are the instructions people were given:

> *This is an experiment to test people's reasoning ability. You will be given eight problems. On each page, you will be shown two statements and you are asked if certain conclusions (given below the statements) may be logically deduced from them. You should answer this question on the assumption that the two statements are, in fact, true.*
>
> *If you judge that the conclusion necessarily follows from the statements, you should answer "yes"; otherwise, "no." Please take your time and be sure that you have the right answer before doing so.*

Here are examples of the syllogisms they were given to consider. Try your hand at them. Remember: Assume the premises are true, and ask yourself whether or not the conclusion must therefore be true as well.

No police dogs are vicious.
Some highly trained dogs are vicious.
Therefore, some highly trained dogs are not police dogs.

No nutritional things are inexpensive.
Some vitamin tablets are inexpensive.
Therefore, some vitamin tablets are not nutritional.

No addictive things are inexpensive.
Some cigarettes are inexpensive.
Therefore, some addictive things are not cigarettes.

No millionaires are hard workers.
Some rich people are hard workers.
Therefore, some millionaires are not rich people.

And the answer is … the first two are valid, but the second two are not.

If you found the answers surprising, that is probably because you succumbed to belief bias: You weighed your prior beliefs more heavily than logical form when evaluating the syllogisms. The conclusions of the first and third syllogism are believable. The conclusions of the second and fourth are unbelievable.

The people in this study showed strong evidence of belief bias. If they had evaluated the syllogisms solely on the basis of the logical form (logical connectedness of premises and conclusion), the acceptance rate for the valid syllogisms would have been 100%, and the acceptance rate for invalid syllogisms 0%. But this isn't what happened. Instead, people seemed to take both believability and logical form into consideration when making their decisions, frequently leading them to make wrong decisions. Valid syllogisms with believable conclusions were correctly accepted about 85% of the time, but those with unbelievable conclusions were accepted only about 55% of the time. Accuracy substantially declined when validity and belief evaluation yielded conflicting decisions! Similarly, invalid syllogisms with unbelievable conclusions were accepted only about 10% of the time, but those with believable conclusions were accepted about 70% of the time! This again was a significant reduction in accuracy.

People are far more likely to succumb to belief bias when they must reason under time constraints. Evans and Curtis-Holmes (2005) allowed people to take as much time as they wanted to evaluate syllogisms like this or restricted their decision time to 10 seconds. When decision time was restricted, valid syllogisms with unbelievable conclusions were accepted less than 40% of the time, whereas invalid syllogisms with believable conclusions were accepted nearly 80% of the time! When rushed, people were far more likely to rely simply on their prior beliefs when evaluating arguments rather than analyzing the logical connectedness of the statements.

In the previous chapters, we saw that decision making is subsumed by two neurologically separate pathways. In the cognitive-science literature, these were referred to as system 1 and system 2. *Dual-process models*, as they are called in cognitive science, have been invoked to explain a variety of cognitive phenomenon. System 1 outputs are fast, automatic, (often) emotion-laden, and take place outside of conscious awareness. When you have a gut instinct about something, or a fact pops into your mind suddenly for reasons unknown, you have experienced the output of system 1. In contrast, system 2 is a slower, controlled, and less emotional process that relies heavily on conscious mental activity to reach a decision. When the outputs of these separate systems agree, a relatively rapid decision can be reached with a high degree of confidence. When they yield different decisional outputs, however, conflict resolution must be invoked to override one or the other, slowing the decisional process considerably. Belief bias shows that sometimes system 1 overrides system 2 outputs.

This theoretical division of reasoning into two separate systems is supported by a good deal of neuroscience research in which people evaluated syllogisms and arguments while the electrical activity on their scalps was measured (Event Related Potentials, or ERP) and when they underwent fMRI imaging of their brains (Goel & Dolan, 2003; Luo and colleagues, 2008). More frontal-lobe activity occurs when people engage in reasoning than when they are engaged in non-reasoning tasks. But different areas of the frontal lobes are active when they make correct and incorrect judgments. The anterior cingulate cortex becomes very active when people are processing a syllogism or argument that pits belief against logical form, registering a conflict between the systems. When they succumb to belief bias, the ventromedial (toward the middle) of the prefrontal cortex is active; but when they make logic-based decisions, the dorsolateral (toward the back and sides) of the prefrontal cortex is active.

How can we become more logical? It turns out that training in disciplines that make heavy use of symbols and symbol manipulation improves people's ability to recognize and evaluate logical form. In one study, Inglis and Simpson (2007) gave simple first-order arguments that differed in believability to people to evaluate for validity. There were two groups, first-year undergraduate students from a high-ranking United Kingdom university mathematics department and trainee teachers not specializing in mathematics. The results showed that the math students were six times less likely than the trainee teachers to succumb to belief bias!

What to Do When the World (or Your Mind) Changes

More often than not in real life, we encounter arguments that are extremely compelling, but not deductively valid. We described these earlier as inductive arguments. But now we must take this one step further. Often, the relationship of support between premises and conclusion is a tentative one, potentially *defeated* by additional information. Under these circumstances, an intelligent reasoner must be prepared to withdraw conclusions in the face of contrary information. This is called *defeasible reasoning* (Pollock, 1987).

In the field of artificial intelligence, this distinction is sometimes referred to as a distinction between monotonic and non-monotonic reasoning (McCarthy, 1980). In monotonic reasoning, inferences are made on the basis of new inputs, current beliefs, and rules. These inferences constitute true beliefs that remain in the knowledge base – so knowledge "grows "monotonically. In *non-monotonic reasoning*, inferences and beliefs can be proven false in light of new information, so they are removed from the knowledge base. As a result, the knowledge base grows and shrinks dynamically.

Consider this situation described by Pollock (1987). You see what appears to be a red cloth. So you reason something like this:

The cloth looks red.
Therefore, the cloth is red.

But then you discover that the cloth is illuminated by a red light. So you modify your conclusion:

The cloth is illuminated by a red light.
Therefore, the cloth may or may not be red.

Pollock uses the term *defeaters* to refer to information that rebuts conclusions outright or undermines the inferential link between premise and conclusions. Both prior beliefs and new information can act as defeaters. When people allow prior beliefs to override logical form, they are sometimes engaged in defeasible reasoning.

For example, consider again the causal argument that appeared to be a case of Modus Ponens:

If the brake is pressed, then the car will slow down.
The brake pedal is pressed.
Therefore, the car will slow down.

Cummins and colleagues (1991, 1995, 1997) showed that people are far less likely to accept the conclusions of arguments like that than ones like this:

If she touches the glass with her bare fingertips,
then her fingerprints will be on the glass.
She touches the glass with her bare fingertips.
Therefore, her fingerprints will be on the glass.

Why? Because people can think of many defeaters for the brake-slow scenario but not for the touch-fingerprint scenario. There are many ways a car could slow down other than having its brake pressed (alternative causes). The car could run out of gas, or it could be going up hill, or the terrain could be getting rough, and so on. There also are many ways a car could fail to slow down even though the brake was pressed (which Cummins referred to as disablers). The brake lines could be cut, the brake fluid may have run out, the road may be very slippery, and so on. This is a very defeasible argument because it admits many alternative causes and many disablers.

But consider the fingerprints argument. It is extremely difficult to come up with an alternative cause for someone's fingerprints being on a glass other than their touching it, and it is extremely difficult to think of disablers that would prevent her fingerprints from ending up there if she touched it. Using Pollock's terminology, the first argument is compelling but defeasible; the second is simply deductively valid.

Casting reasoning in this light makes it clear that when people reason, they are trying mostly to maintain truth and consistency in their knowledge bases. Retrieving facts from memory is easier and faster than extracting and evaluating logical form, as the TV program *Jeopardy* shows us daily. When hard-won beliefs are pitted against logical form, beliefs frequently hold sway, particularly if we are rushed. Furthermore, prior beliefs can often legitimately undermine what appears to be a deductively valid argument.

As we'll find out in subsequent chapters, this reliance on prior beliefs is both a blessing and a curse. Our beliefs can guide us away from making bad choices. But they can also make us pigheaded.

Box 5.1. How Aristotle Thought

A *categorical syllogism* is an argument consisting of exactly three categorical propositions (two premises and a conclusion) that uses exactly three categorical terms, each of which is used exactly twice. Here is an example:

All geese are birds.
All birds have feathers.
Therefore, all geese have feathers.

Categorical syllogisms were introduced by Aristotle and formed the cornerstone of his logical system of reasoning. Medieval logicians devised a simple way of labeling the various forms in which a categorical syllogism may occur. The argument above has the form AAA-1, and it is deductively valid. Here it is stripped down to its essentials:

All A are B.
All B are C
Therefore, all A are C.

No matter what A, B, or C stand for, an argument of this form is deductively valid. Here's another:

No P are M.
Some S are not M.
Therefore, some S are not P.

This is an invalid argument of the form EOO-2. For example

No dogs are cats.
Some birds are not cats.
Therefore, some birds are not dogs.

Sure, some birds are not dogs. In fact, all birds are not dogs. But you can't get there from the stated premises. The premises don't connect birds to dogs in any logical way. They just tell you that dogs and cats are different kinds of animals, and that some birds are different from cats. That doesn't mean they are different from dogs. So the premises don't guarantee the truth of the conclusion.

This isn't a textbook on logic, so you really don't need to fully understand categorical syllogisms or Aristotelian logic. What you do need to appreciate, though, is that *logical validity depends solely on the form of the argument*. The content of the propositions is irrelevant.

Aristotelian logic stood alone as a means of capturing inference for two thousand years. Historian of logic Karl von Prantl (1820–1888) went so far as to claim that that any logician after Aristotle who said anything new was confused, stupid, or perverse (*Stanford Encyclopedia of Philosophy*). But the difficulty is that Aristotelian logic severely limits what can be expressed and what can be argued and proven. More powerful means of capturing the expressive power of natural language and natural inference were needed, and, in the nineteenth and twentieth centuries, many answered the call.

In the mid-nineteenth century, George Boole (1815–1864) developed a mathematical-style "algebra" that extended Aristotelian logic by permitting

(*continued*)

an argument to have many premises and to involve many classes. This algebraic approach was later rejected by Alfred Whitehead and Bertrand Russell in favor of an approach developed by Gottlob Frege that made use of logical connectives, relation symbols, and quantifiers. Their goal was an ambitious one, namely, "to show that all pure mathematics follows from purely logical premises and uses only concepts definable in logical terms" (Russell, 1959, p. 74). The culmination of their work was the publication of the three-volume masterpiece *Principia Mathematica* (1910–1913). Unfortunately, this worthy goal was finally proved impossible when Gödel showed that even all the truths of arithmetic cannot be deduced from any set of premises.

SIX

What Causes What?

At bottom, we all have a little bit of the control freak in us. We want to know how to bring about the things and events that we like and how to prevent or terminate the things and events we don't like. Some of us also just want to know what causes what even if there is nothing we can do about it. We strive for that state of cognitive satisfaction that goes something like this: "Huh! So that's why that happens." In order to accomplish these goals, we automatically rely on a very basic concept, one that philosopher John Mackie (1974) dubbed "the cement of the universe": *causality.*

The Paradox of Causality

From a psychological viewpoint, causality is a paradox. We use this concept to make sense of events in our everyday lives as wide ranging as why our car didn't start to why people commit acts of violence. Yet it eludes the senses – it cannot be directly perceived – which led philosopher David Hume (1711–1776) to claim that causality was an "illusion" (Hume, 1748). As he pointed out, if one event causes another, then (a) the two events always co-occur (constant conjunction), (b) they must occur close in time, with the cause preceding the effect (temporal priority), (c) they must occur close in space (spatial proximity), and (d) one event must have the power to bring the other about (necessary connection). Now the paradox: Whereas we can directly perceive the first three, we cannot see the fourth – a necessary connection between events. As a simple example, imagine watching a sledgehammer hit a crystal vase, followed by the vase shattering. Read that sentence again. It contains all the information about the event that is directly perceivable. So you should have something like this in your head: "The hammer hit the vase, *and then* it shattered." But that is not what is in your mind right now. In your mind is a proposition like this: "The hammer hit the vase, *causing* it to shatter." Where did that term "causing" come from? Causation cannot be directly perceived. That is why Hume claimed that causation is an illusion imposed by the mind, in much the same way that our visual system is subject to various illusions. This is how he put it: "Power and Necessity exist in the mind, not in objects [They] are consequently qualities of perceptions, not of objects, and are internally felt by the soul, and not perceiv'd externally in bodies" (Hume, 1748, p. 166). Hume saw no other way of explaining the concept of causality because he was one of the founders of British Empiricism, a philosophical position that holds that all knowledge is derived from sensory experience. We are born as blank slates; anything we know, we've learned through sensory experience. The concept of causation is problematic for such a position because a causal connection cannot be directly sensed or perceived. So it must be an illusion in much the same way that seeing a mirage in the desert is an illusion.

The German philosopher Immanuel Kant (1724–1804) violently objected to this view of causality. As he put it, Hume's position made the concept of *cause* "a bastard of the imagination, impregnated by experience" (Kant, 1783, p. 258). He argued instead that the existence of causality is an a priori truth – a truth that is knowable through reason alone. In Kant's view, we are incapable of experiencing or thinking about an a-causal world because the concept of causality is implicit in the form of our judgments. To put it another way, our judgments have certain forms, and the category or concept of causality is implicit in one of these forms. If you couldn't make judgments of this form, you wouldn't be a normal rational agent. So causality is not an illusion; it is knowledge. From a twenty-first-century standpoint, we can almost paraphrase this in the following way: Causality is an innate category of knowledge that we apply to the world when interpreting certain kinds of events. It is a natural part of our cognitive architecture.

How the Experts Decide What Causes What

These ideas from eighteenth-century philosophers are very much alive and well in current psychological literature on causal cognition. One very influential contemporary treatment of causal cognition was proposed by Patricia Cheng (1997) and Laura Novick (Cheng & Novick, 1992). As Cheng puts it, "Causal relations are neither directly observable nor deducible … the reasoner believes that there are such things in the world as causes that have the power to produce an effect and causes that have the power to prevent an effect, and that only such things influence the occurrence of an effect" (Cheng, 1997, p. 372). From a psychological standpoint, people interpret events in the world in causal terms. But for any given event, there are a multitude of possible explanations (causes). When your car doesn't start, it could be that your battery is dead, there is no gas in the car, or any number of other reasons. How do we determine the true cause of the event?

According to Cheng and Novick, reasoners evaluate *covariation* information in order to select among possible explanations. The one with the strongest statistical contingency (or covariation) is selected as the cause of the event. This can be captured formally by the following model:

$$\Delta P = p(e|c) - p(e|{\sim}c)$$

e stands for the effect
c stands for a candidate cause
$p(e|c)$ is the probability of e given the presence of c
$p(e|{\sim}c)$ is the probability of e given the absence of c

In English, this equation calculates the difference between the probability of the effect occurring in the presence of the cause and in the absence of the cause. As Hume pointed out, causes and effects must co-occur. Going back to our car example, suppose you just bought a used car, and, unbeknownst to you, there is a malfunction that drains the battery if you play the radio for too long. So if you've listened to the radio while driving, you frequently find that you can't start your car again because the battery's dead. ΔP would be very high in such a case because the co-occurrence of playing the radio and having a dead battery is greater than the co-occurrence of not playing the radio and having a dead battery. You may even utter in exasperation, “It seems like practically every time I listen to the radio, the battery goes flat!”

According to Cheng and Novick, if ΔP is noticeably positive (above some criterion), then *c* is considered a generative or facilitory cause. If it is negative, then *c* is an inhibitory or preventive cause. If it is below criterion, *c* is judged to be non-causal. Cheng (1997) took this analysis one step further by proposing a model that better distinguishes between causal events and coincidences:

$$p_c = \Delta P/[1{-}p(e|{\sim}c)]$$

This model compares the causal strength of a purported cause to the likelihood of the effect occurring in the presence of other causes. Let’s plug in some numbers and see how it works. Suppose you’ve listened to the radio on ten occasions and didn’t listen to the radio on another ten occasions. Eight of the times you listened to the radio, the battery died. So the probability that the battery dies given that you’ve listened to the radio is 8/10 = .8. On 80% of the occasions when you listened to the radio, the battery went flat. Now what about the times you drove without playing the radio? Suppose on two out of those ten occasions, the battery went dead anyway. The probability that the battery goes flat given that you did not listen to the radio is 2/10 =.2 (or 20%). Calculating ΔP is a piece of cake: .8 – .2 = .6. Given that ΔP ranges from –1 to 1, with 0 meaning no causal connection, a value of .6 certainly suggests that there is a facilitative causal connection (playing the radio causes the battery to go flat), but it isn’t particularly strong evidence. Let’s take a look at the occasions you suffered a dead battery but weren’t listening to the radio. This means there are alternative causes (such as a fault in the battery itself). The denominator in the above equation expresses the likelihood that the effect was caused by the radio rather than an alternative cause. The expression p(e|~c) is the probability of the effect occurring without the cause – that is, a dead battery even though the radio was not played. We calculated this before: 2 out of 10 times or .2. Subtracting this from 1 gives us the proportion of times the effect was present when the cause did occur. Also easy to figure: 1 – .2 = .8. If we now divide ΔP by this value, we will get $p_{c,}$ which will tell us how much power can be attributed to the connection between the radio and the dead battery compared to other possible causes. In our example, this value turns out to be .6/.8 = .75. Given that p_c can range from 0 to 1, this is stronger evidence that the radio is the culprit. So, in our experience, the radio was moderately (.6) implicated as a cause of our car's dead battery, but compared to the alternatives, it is a very likely candidate (.75).[1]

Now the fly in the ointment: Cheng and Novick's models tell us how to choose among potential explanations for an event using covariation information. But it turns out to be only part of the story.

Here is a simple way to illustrate the problem. Using the same numbers as above, imagine that, instead of playing the radio, eight out of the ten times your battery went dead, you were taking a turkey sandwich to work for lunch. Our values of ΔP and p_c would be exactly the same. Yet you would find it very hard to believe that carrying a turkey sandwich could cause the battery to go flat. Cheng accommodates for this by saying that turkey sandwiches would not be included in your focal set of possible causal candidates. But this begs the question why.

Causal-power theorists take this objection very seriously. Their main claim is that people make causal judgments by seeking information about (or inferring) possible generative causal mechanisms. As Ahn and colleagues (1995) put it, "… a mechanism is some component of an event which is thought to have causal force or causal necessity.… Underlying two causally linked events, there is a system of connected parts that operate or interact to make or force an outcome to occur." To demonstrate this, they presented people with descriptions of causal events, such as "Kim had a traffic accident last night" followed by statements that made reference either to causal mechanism (e.g., "Kim is nearsighted and tends not to wear her glasses while driving") or

[1] Mathematician Judea Pearl (2000) has shown that Cheng's causal-power theory can be given a counterfactual interpretation (i.e., the probability that, absent *c* and *e*, *e* would be true if *c* were true). The implication is that Cheng's model is computable using structural models. Griffiths and Tenenbaum (2005) further showed that the model can be interpreted as a noisy-OR (disjunction) function used to compute likelihoods.

covariation information (e.g., "Traffic accidents were much more likely last night"). The participants were instructed to rate to what degree each factor was responsible for the event. The ratings given showed that mechanism information was considered roughly twice as effective as covariation information in producing the effect. In another experiment, participants were told to write down any questions they would want answered in order to identify the causes of the events. The results were quite striking: People rarely requested covariation information when asked to determine the cause of an event, even when such information was readily available. Instead, their requests were theory-based, aimed at uncovering a mechanism that could bring the event about. Even when covariation information was requested and given, reasoners did not justify their attributions through reference to covariation information. Instead, the majority of causal explanations subjects offered were mechanism-based rather than covariation-based. Further, White (1995, 2000) found that even when a causal candidate was perfectly correlated with an effect, it was not identified as a cause unless it was also believed to possess the power to produce the effect. It seems people insist on that "necessary connection" that Hume found so problematic.

To make this clearer, consider this simple example: Suppose the incidence between car accidents perfectly correlated with (a) drivers having tatoos and (b) a new braking mechanism. People will opt for (b) as the true cause and flatly refuse to believe (a) – unless they believe tatoos covary with something else that could have generative causality (such as people who have tatoos are more likely to drive while under the influence of drugs or alcohol, or that the tatoos release toxic substances into the body that impair judgment).

People are not far wrong in their skepticism of relying too greatly on covariation in drawing causal inferences. Judea Pearl (2009), a computer scientist and philosopher who is a leading expert on causal modeling, argues that it is futile to try to define causality purely in probabilistic terms. He argues that cases that appear to do so, on closer inspection, typically rest on hidden causal assumptions that are best captured through counterfactual ("if this hadn't happened, that wouldn't have happened either") or mechanistic assumptions. Unfortunately, this sometimes leads to biases in human causal reasoning whereby people simply dismiss objective data supporting a causal theory if they cannot understand the theory or they find the theory objectionable. A contemporary example of this is the refusal of a majority of Americans to accept the theory of evolution, despite ample evidence supporting its validity; according to a 2009 Gallup Poll, only four in ten Americans accept the theory of evolution.

How Your Brain Decides What Causes What

We can think of theory-based causal reasoning as belief-based reasoning. We have prior beliefs concerning the underlying mechanisms that could allow one event to bring about another. The impact of prior beliefs on causal cognition was dramatically demonstrated in a neuroscience study conducted by Fugelsang and Dunbar (2005) in which participants evaluated causal scenarios while undergoing fMRI. Participants were shown four sets of materials, two that described plausible causal scenarios and two that described implausible causal scenarios. In the plausible conditions, they were told that higher levels of serotonin improves mood (plausible), and a red pill was identified as a drug that increases serotonin levels. In the implausible conditions, they were told that antibiotics have no impact on mood, and a red pill was identified as an antibiotic. They were then shown a randomized series of drawings that paired the red pill with a happy or sad face and a blue pill (placebo) paired with a happy or sad face. The rate of covariation was either 18 out of 22 times, so $\Delta P = .74$ (a moderate to strong causal connection), or 10 out of 22 times, yielding a ΔP of .30 (a weak causal connection). Finally, they were asked to rate how effective the red pill was at increasing feelings of happiness, using a three-point scale (low, medium, and high).

The results were surprising: Areas associated with learning were most active when theory and data were consistent (plausible and strong covariation, or implausible and weak covariation), especially when people were evaluating data that were consistent with a plausible theory. Areas associated with thinking and attention were most active when data and theory were inconsistent (implausible theory with strong covariation or plausible theory with weak covariation), especially when plausible theories had weak support. What does this mean? People devote more attention to processing data that are inconsistent with one's beliefs, but they do not necessarily learn from that process (revise their beliefs). The authors concluded that there is a strong belief bias in causal reasoning that goes like this: (1) Focus on theories that are consistent with your belief, (2) attend to inconsistent data, (3) but do not necessarily revise your beliefs. So where causality is concerned, we are conservative reasoners. We resist changing our minds even in the face of disconfirming evidence.

What Is Necessary? What Is Sufficient?

These results demonstrate that causal reasoning typically involves retrieval and use of prior knowledge or beliefs. So we can ask a slightly different question: How do prior beliefs affect causal inference? Consider, for example, the statement "jumping into a pool full of water causes a person's clothes to get wet." This is a reasonable belief, adopted on the basis of lots of covariation evidence. We know that jumping into a pool full of water does indeed have the causal power to make a person's clothes wet. Now suppose you have been invited to a pool party, and as you approach the front door of the house, a person whose clothes are wet exits. You, having gone to pool parties before, remember that there is usually a lot of splashing while jumping into the pool, general horsing around in the water, and the like. Frequently, this means that people standing by the pool get pretty wet. Do we conclude that the person we just saw has been in the pool?

Under these circumstances, we probably would not draw that conclusion because we believe (we retrieve from memory) many causes for the person's clothes to be wet other than jumping into the pool. Because there are alternatives, we don't think that jumping into the pool is a *necessary cause* for the person's clothes to be wet. Now if we saw this fully clothed person jump into the pool and emerge bone dry, we would be very surprised because jumping into a pool full of water certainly seems *causally sufficient* to produce the effect of wet clothes.

Contrast that situation with the following: You and a friend go canoeing and camping. Later, you watch as she attempts to build a campfire. After piling up leaves, twigs, and dry grass, she strikes a match to light the fire. You would not find it surprising if the match did not light. Normally, striking a match causes it to light. You don't doubt the causal relationship – striking a match does in fact have the power to cause it to light. But there are factors that could intervene that prevent the effect from occurring in the presence of the cause. For example, the matches might be wet from the canoe trip. This doesn't mean that you've disconfirmed the causal relationship between striking and lighting. Instead, it means that striking the match wasn't sufficient to light it. The cause-effect relationship was disabled by the water on the matches.

Notice that two new terms have been introduced: *causal necessity* and *causal sufficiency*. These terms have a long history in philosophy. British Empiricist David Hume (1748, Section VII) offered the following definition of causation: "We may define a cause to be an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second. Or, in other words, where, if the first object had not been, the second never had existed." By this definition, a cause is a necessary condition for the occurrence of a particular event. Notice that Hume expresses causal necessity as counterfactual – if the cause had not occurred, the effect would not have, either. Contemporary philosopher of science David Lewis (1973, 1979) formalized this counterfactual claim, proposing a modal logic to capture its meaning and implications. Echoing Hume, he argued that "where *C* and *E* are actual events, to say that *E* is causally dependent on *C* is just to say that if *C* had not occurred, then *E* would not have occurred." This can be formalized in the following way using modal logic where the box symbol represents necessity: Let *c* and *e* be two distinct possible particular events. Let O(e) represent all and only those possible worlds where *e* (the effect) occurs, and O(c) represent all and only those possible worlds where *c* (the cause) occurs. Then *e* depends causally on c if the counterfactuals shown in Figure 6.1 are true.

Why does Lewis qualify this analysis with the clause "where *C* and *E* are actual events?" Because Lewis argues that non-events (absences) cannot serve as causes. For example, suppose you observe someone climbing a ladder to the roof of a house. The sentence "Climbing the

$O\ (c)\ \Box\!\longrightarrow O\ (e)$ if *c* occurs, then it is necessarily the case that *e* occurs.

$\sim O\ (c)\ \Box\!\longrightarrow \sim O\ (e)$ if *c* does *not* occur, then it is necessarily the case that *e* does *not* occur.

FIGURE 6.1. Philosopher David Lewis's counterfactual treatment of causality.

ladder caused Joe to reach the roof" describes two actual events, and the occurrence of one (Joe reaching the roof) was dependent on the other (Joe climbing the ladder). The sentence "The ladder not breaking caused Joe to reach the roof" is not a causal claim (despite the dependency described) because the ladder not breaking is a non-event. The fact that people often attribute causal status to non-events eventually led Lewis (2000, 2004) to the more extreme claim that causation is not a relation between events.

British Empiricist John Stuart Mill (1843) offered an explicit analysis of the terms causal necessity and causal sufficiency. According to Mill, a necessary cause of *e* is a factor that is present in all cases of an effect *e*. A sufficient cause of *e* is a factor that guarantees the existence of effect *e*. A factor can be singular events (the battery going flat), properties (the battery having a full charge) or variables (temperature). The factor that made the biggest difference to *e*'s occurring is the *central factor* (e.g., the battery going flat). Finally, Mill proposed the *deterministic principle*, which states that the same causal antecedents always produce the same effects (e.g., batteries going flat always lead to cars not starting). Together, these claims are referred to as *Mill's Canons*.

Philosopher John Mackie (1974) clarified these concepts further by offering the following analysis:

- Sufficiency means that a cause can, by itself, produce an effect. *(If c, then e.)*
- Necessity means that a particular cause must be present for an effect to occur.
- *(If not c, then not e.)*
- A necessary and sufficient causal relation is one in which there is only one cause for an effect. *(If and only if c, then e.)*

His most famous contribution is his INUS theory: A cause is an *in*sufficient but *n*ecessary part of a scenario that is *un*necessary but *s*ufficient for an event to take place.

The importance of distinguishing between causal necessity and sufficiency is perhaps most apparent in scientific explanation. Philosophers Nancy Cartwright (1980) and David Armstrong (1983) have argued that the actual laws of nature are *oaken* rather than *iron*. *Oaken* laws admit exceptions: They have tacit *ceteris paribus* (other things being equal) or *ceteris absentibus* (other things being absent) conditions. For this reason, an inference based on a law of nature is always defeasible, since we may discover that additional factors must be added to the law in question in special cases.

How What You Believe Influences How You Decide

As discussed in the previous chapter, consideration of exception information is strongly implicated in everyday human reasoning. People use this type of information to make judgments concerning causal necessity and sufficiency to guide their causal judgments.

Specifically, Cummins and colleagues have shown that when people evaluate causal arguments, they activate and retrieve information from memory concerning alternative causes and disablers (Cummins, 1995, 1997; Cummins et al., 1991). More specifically, they consider alternative causes when deciding whether an event will occur or whether to attribute a causal role to a particular event; they also consider whether possible "disablers" can prevent an event from occurring even though a plausible cause is present. Alternative causes cast doubt on causal necessity, and disablers cast doubt on causal sufficiency.

Disablers are of crucial importance in our causal reasoning because they constitute (or suggest) interventions for preventing unwanted events. For example, it is widely known that being abused as a child makes an individual more likely to be an abusive parent. But the relationship is anything but certain. In fact, research indicates that numerous factors can produce abusive child-rearing practices, but the presence of a non-abusive role model during childhood reduces the likelihood that an abused child will become an abusive parent (Kaufman & Zigler, 1988; Martin & Elmer, 1992). Non-abusive role models, therefore, are disablers who can prevent an effect (future abusive parenting) from occurring despite the presence of a true cause (being abused). This suggests an important way to intervene if one suspects child abuse, particularly if it is not possible to stop the abuse itself.

Disablers do not nullify a causal relationship. Child abuse remains a true cause of later abusive parenting, but the cause-effect relationship is not inevitable. It can be interrupted or disabled by mitigating factors, such as the presence of a non-abusive role model. This is the reverse of *enabling conditions* – background factors that must be present for a true cause to produce an effect, although they themselves are not causes. An example is oxygen, which enables combustion when combined with a true cause, such as striking a match. Disablers are background factors that must be absent in order for a true cause to bring about an event. Using Cheng's terminology, we can think of them as preventive causes (although I have always found that term awkward).

Activation of one's knowledge of alternative causes and disablers constitutes a core part of the causal-reasoning process. These factors impact reasoning judgments even when they are not explicitly mentioned in the reasoning scenario. Cummins and colleagues found that people were reluctant to conclude that a particular cause produced an effect if many alternative causes were possible, and they were reluctant to conclude that a cause was sufficient to produce an effect if many disablers were possible. Thus, in everyday reasoning, alternative causes do indeed cast doubt on causal *necessity* whereas disablers cast doubt on causal *sufficiency*. This effect has been replicated many times with adults (e.g., de Neys et al., 2002, 2003) and children (e.g., Janveau-Brennan & Markovits, 1999).

Further, Verschueren et al. (2004) had people think aloud while making causal judgments and found that people spontaneously retrieve alternative causes when deciding whether a particular cause is necessary to produce an effect, and they spontaneously retrieve disablers when deciding whether a particular cause is sufficient to produce an effect. The greater the number of items retrieved, the slower reasoners are to come to a decision (de Neys and colleagues, 2003).

The Causal Paradox Revisited: What Infants Told Us

I'd like to end this chapter by returning to the paradox raised at the beginning: How is it that we can think about causality if we cannot directly perceive it? One answer to this paradox comes from research on early emerging knowledge in infancy. The bottom line of developmental research over the past several decades is that much of the core knowledge we need to make sense of the world is either present at birth or emerges very quickly in early childhood – too quickly for it to have been laboriously learned from trial-and-error experience. One aspect of this core knowledge is the concept of causality, which emerges within the first six to eight months of life.

The earliest experiments done to investigate the emergence of causal knowledge in infants made use of a methodology pioneered by Albert Michotte, a Belgian experimental psychologist. According to Michotte (1963), causality is perceived in terms of a transfer of motion, energy, momentum, or force. The simplest demonstration of these concepts are causal motion events, such as launching, where one object collides with another and causes it to move. He believed the capacity to represent or perceive such events as causal is innate and constitutes the foundation of all later developing causal knowledge.

Michotte did not test his ideas on infants, but he did devise and test a number of demonstrations of these principles with adults. The design is simple yet powerful: The participant watches as two simple objects (such as a red ball and a blue ball) interact with each other in various ways on a projection screen. In the "launching with direct contact" display, for example, the blue ball sweeps across the screen and hits a stationary red ball, which then moves off in the same direction. When observing this event, most adults perceive it as a causal *launching* event – that is, the blue ball hits the red ball, thereby causing the red ball to move. In a *delay* condition, the blue ball sweeps across the screen and makes contact with the red ball, but the red ball moves only after a brief delay. Most adults do not perceive this as a causal event. In a *spatial-gap* display, the blue ball sweeps across the screen but stops short of touching the red ball. The red ball then moves off. Adults generally do not see this as causal either, presumably because there is no contact between the two objects. Instead, it looks as though the red ball moved of its own accord. The important aspect of this for our purposes is that Michotte believed that perception of motion events based on spatiotemporal input alone was the only ontogenetic source of causal reasoning.

About ten years later, developmental psychologists began testing Michotte's theoretical view by showing infants of various ages these kinds of displays and recording their attention using a habituation paradigm (e.g., Ball, 1973). This methodology is quite simple: Infants are shown an event repeatedly until they become bored and look away (such that their viewing time is 50% less than when the display was initially presented). Then the display is changed in some theoretically relevant way, and the amount of time the infants spend looking at the new display is recorded. If they recover interest (viewing time increases significantly), they can tell the difference between the old and new display. Using launching displays as examples, the infants would be shown a blue ball sweeping repeatedly across a screen until they become bored, reducing their viewing time by 50%. Then they would be shown the direct-launching, delayed-launching, or gap-launching display. Since these are all new displays as far as they're concerned, they should recover interest in each case. Yet they don't. Instead, they find the latter two far more interesting then the direct launching. This suggests that they already know about causality, so the display that obeys the laws of causality is far less interesting than the displays that seem to violate it. This extends to cases where the blue ball collides with the red ball but the red ball does not move; they find it surprising when the cause occurs but the effect does not (Kotovsky & Baillargeon, 2000).

In some studies, the actual launching event is hidden (occluded) partially behind a screen. For example, the display shows a blue ball, a screen, and a red ball that is half-hidden behind the edge of the screen. The blue ball comes racing along, passes behind a screen, and then the red ball comes racing out from behind the screen. In this kind of setup, infants are prevented from seeing the causal event, so they must infer that the blue ball makes contact with the red ball. And they do, recovering interest more if the screen is lifted to show the blue ball is not hitting the red ball than when it is hitting the red ball.

The interesting question is when do infants start to show this kind of knowledge? Convergent evidence from many sources suggests that by six to eight months of age, infants perceive causality in Michottian launching events (see Muentener & Carey, 2010, for a comprehensive review of this literature). But certain aspects of causality are not present yet at this age. For example, although adults certainly notice the difference between temporal and spatial-gap launching displays, infants don't seem to. If they are habituated to one, they do not recover interest when the display is changed to the other. Appreciation of these and other aspects of causal events emerge later in infancy.

These results perhaps offer a solution to the dispute between Hume and Kant. Causality is indeed a property of physical events, and we do indeed interpret events in causal terms. But the results of careful investigation of infant cognition suggest that this is more than just an illusion. The systematicity in infants' early discrimination of causal and non-causal events and the greater refinement of their discrimination of subclasses of causal events in late infancy suggest instead that this constitutes knowledge, not perceptual illusion.

So what have we learned about human causal reasoning? We have learned that it can be summarized by two key points. The first is that people are sensitive to covariation between events, and often base their causal inductions on the strength of the covariation. The second is that we also need a plausible story connecting the two events. If we've got both we believe we've got a case of full-blown causality – even if the plausible story happens not to be true or is not adequately validated.

To put it another way, we are plainly biased toward inferring (or perceiving) causality in the face of reliable covariation, even when we should not be. That means we are biased toward making false-positive errors rather than false-negative errors. And there may be good reason for that: False-negative errors can have more devastating survival consequences than false-positives.

See that dark spot in the bushes? Is that a predator lurking, or a simple shadow? Making a false-positive error means concluding that the dark spot really is a predator (when it is not) and fleeing the area.

Making a false-negative error means concluding it is nothing important (when it is a predator) and staying put. Better to err on the side of caution. Feeling nauseous? Is that because of the funny-tasting food you just ate or is it just a coincidence? A false-positive error means inferring that the funny-tasting food made you sick (when it did not) and avoiding it in the future. A false-negative error means concluding it was just a coincidence (when it did make you sick) and continuing to eat it.

In terms of survival, you get more bang for the buck by allowing more false-positives than false-negatives. In terms of causality, this means it is beneficial to have a bias toward inferring causality in the face of reliable covariation ("this food made me sick") than neutrality ("my nausea was probably just a coincidence").

The problem, of course, is that we want to maintain a knowledge base filled with *true* beliefs, particularly true beliefs about what causes what. Typically, this means we have to test the truth of our beliefs via a process called hypothesis testing: assume the belief is true, then deduce what it predicts, and test those predictions. As we will see in the next chapter, people are genuinely bad at this. If you have not yet learned how to test the truth of hypotheses (or any other kind of statement), the next chapter is the most important chapter you will ever read.

SEVEN

Hypothesis Testing

TRUTH AND EVIDENCE

In 1960, British research psychologist Peter Wason reported a curious phenomenon in human reasoning. He gave his participants this simple task:

> You will be given three numbers that conform to a simple rule I have in mind. This rule is concerned with a relation between any three numbers and not with their absolute magnitude, i.e., it is not a rule like all numbers greater (or less) than 50, etc.
>
> Your aim is to discover this rule by writing down sets of three numbers, together with reasons for your choice. After you have written down each set, I shall tell you whether your numbers conform to the rule or not, and you can make a note of this outcome on the record sheet provided. There is no time limit, but you should try to discover this rule by citing the minimum number of sets of numbers.
>
> Remember that your aim is not simply to find numbers that conform to the rule but to discover the rule itself. When you feel highly confident that you have discovered it, and not before, you are to write it down and tell me what it is.
>
> Here are the numbers that conform to the rule I have in mind: **2 4 6**

Before you continue, try to guess what the rule is. Then write down some triples that would help you find out whether or not you're right. That is, if you showed the triples to Wason, he could tell you whether or not each one was an instance of the rule he had in mind.

Confirmation Bias: Tell Me I'm Right

If you are like the majority of Wason's participants (54%), the rule you're entertaining – your hypothesis – is something like "increasing intervals of two" or "consecutive even numbers." But you would be wrong.

Moreover, if you are like Wason's participants, the majority of the triples (77%) that you wrote down were almost all *positive test instances* of your hypothesized rule; they all conform to the rule you have in mind because they have one or more of the properties described by your hypothesis. For example, you may have written "4 6 8," "22 24 26," or "340 342 344." These are all instances that are consistent with the rules "increasing levels of two" and "consecutive even numbers."

Congratulations! You have just demonstrated two very robust predilections of human reasoners: You have chosen to seek evidence that *confirms* your beliefs, and you have chosen to test your hypothesis by generating positive test cases that have a property of interest (e.g., even numbers) rather than those that do not (e.g., odd numbers). But is your hypothesis correct? And was your evidence-seeking strategy optimal?

Wason would have indeed confirmed that each of your triples were instances of the rule he had in mind. Unfortunately, the rule he had in mind is not the one you proposed. So your hypothesis – your belief – is wrong.

Suppose instead that you had chosen to try to *falsify* your belief. You would then have included an instance that clearly violated the rule you had in mind, such as "1 3 8." This is called a *negative test instance* because it includes properties that are not part of your hypothesized rule (i.e., odd numbers).

You probably would have been surprised to find that "1 3 8" does in fact conform to the rule he had in mind. This is a very useful piece of information. It shows that the triples do not need to be consecutive even numbers. In fact, they do not need to be even at all.

Let's try another triple that violates the last remaining remnant of your belief – namely, that the numbers must be ascending. Let's try "6 3 1." Wason would tell you this triple is not an instance of the rule he had in mind. This is also very useful. It would suggest that the series must be ascending. In fact, that is the pure and simple totality of the rule: *any ascending series of numbers.*

This simple demonstration nicely showcases a number of pitfalls in our natural intuitions about hypothesis (or truth) testing. First, it shows that the first set of instances that we encounter may bias our hypotheses by making them overly constraining (or, in some cases, too lenient). Next, it shows that once we develop a belief or hypothesis, we seek evidence that proves we're right rather than seeking evidence that might prove we're wrong. We do this by generating positive instances to test our hypothesis. Third, it shows that people are very resistant to changing the nature of their beliefs in the face of contrary evidence: Even after hearing "no, that is not an instance," many of Wason's participants clung to their hypothesis but changed their triples to confirm some other aspect of it.

A More Realistic Study of Confirmation Bias

Perhaps this study seems a bit precious to you. After all, does it really matter how people go about testing the truth of arbitrary mathematical rules that reside only in the mind of a research psychologist? But consider the following study by Lord, Ross, and Lepper (1979).

The participants were forty-eight university students. Twenty-four of them were "proponents" who favored capital punishment, believed it to have a deterrent effect, and thought most of the relevant research supported their own beliefs. The remaining twenty-four "opponents" opposed capital punishment, doubted its deterrent effect, and thought that the relevant research supported their views. They were shown a brief statement about a single study on capital punishment such as

> Kroner and Phillips (1977) compared murder rates for the year before and the year after adoption of capital punishment in 14 states. In 11 of the 14 states, murder rates were lower after adoption of the death penalty. This research supports the deterrent effect of the death penalty.

Participants were told the study was factual, although it was in fact fictitious. They were then asked to answer the following two questions using a rating scale from −8 (more opposed) to +8 (more in favor):

> Has this study changed the way you feel toward capital punishment?
> Has this study changed your beliefs about the deterrent efficacy of the death penalty?

Following this, they were given more detailed information about the study, including the researchers' procedures, reiteration of the results, summaries of several prominent criticisms of the study, and the authors' rebuttals to the criticisms. They were then asked to judge how well or poorly the study was conducted on a scale of −8 (very poorly) to +8 (very well), how convincing the study seemed as evidence on the deterrent efficacy of capital punishment (from −8 = completely unconvincing to + 8 = completely convincing), and to write a summary explaining why they thought the study did or did not support the argument that capital punishment is a deterrent to murder.

The entire procedure was then repeated with a second fictitious study reporting results opposite to those of the first:

Palmer and Crandall (1977) compared murder rates in 10 pairs of neighboring states with different capital punishment laws. In 8 of the 10 pairs, murder rates were higher in the state with capital punishment. This research opposes the deterrent effect of the death penalty.

Half of the proponents and half of the opponents saw a "pro" study result first, and the remaining halves saw an "anti" study first.

This is a more realistic reasoning scenario. It does not differ much from the way we generally try to educate students about political issues. It is also how our politicians and lawmakers debate such issues. So what did the researchers find?

The striking results of this study can be simply summarized this way: *The two groups became more certain of their original positions as a result of reading and evaluating the evidence presented, thereby diverging even more than they did initially!*

The researchers concluded that people are likely to examine relevant empirical evidence in a biased manner. This bias expresses itself as a tendency to accept confirming evidence at face value while subjecting disconfirming evidence to critical evaluation. The result, therefore, of exposing people to objective evidence relevant to deeply held beliefs may be not a narrowing of disagreement but rather an increase in polarization. Or to quote Sir Francis Bacon (1620),

> The human understanding when it has once adopted an opinion draws all things else to support and agree with it. And though there be a greater number and weight of instances to be found on the other side, yet these

> it either neglects and despises, or else by some distinction sets aside and rejects, in order that by this great and pernicious predetermination the authority of its former conclusion may remain inviolate.

Box 7.1. Seeing What We Expect to See

If you are still unconvinced that our beliefs influence how we interpret facts, consider the example of the "flying horse" (Olsen, 2004). Depictions of galloping horses from prehistoric times up until the mid-1800s typically showed horses' legs splayed while galloping – that is, the front legs reaching far ahead as the hind legs stretched far behind. People just "knew" that was how horses galloped, and that was how they "saw" them galloping. Cavemen *saw* them this way, Aristotle *saw* them this way, and so did Victorian gentry. But all of that ended when, in 1878, Eadweard Muybridge published a set of twelve pictures he had taken of a galloping horse in the space of less than half a second using twelve cameras hooked to wire triggers. Muybridge's photos showed unequivocally that a horse goes completely airborne in the third step of the gallop with its legs *collected* beneath it, not splayed. It is called the moment of suspension. Now even kids draw horses galloping this way.

When Your Brain Is Biased

Just prior to the 2004 presidential election, a group of researchers conducted an fMRI study involving thirty men, half of whom described themselves as staunch Republicans and half as staunch Democrats (Westin and colleagues, 2006). While being scanned, they were asked to assess contradictory statements by both George W. Bush and John Kerry. They were shown eighteen sets of stimuli, six each regarding President George W. Bush; his challenger, Senator John Kerry; and a politically neutral male control figure (such as actor Tom Hanks). For each set of stimuli, the men first read a statement made by one of these individuals. This statement was followed by evidence that the individual had done something that contradicted his initial statement. When faced with information like this, most people suspect that the person in question is dishonest or pandering to constituents. They say one thing and do another.

Next, the men were asked to rate the extent to which the person's words and deeds were contradictory. Finally, they were presented with a statement that might explain away the apparent contradiction and asked to reconsider and again rate the extent to which the target's words and deeds were contradictory.

The men's verbal responses showed a pattern of emotionally biased reasoning: They denied noticing obvious contradictions for their own candidate that they had no difficulty detecting in the opposing candidate. Republicans and Democrats did not differ in their responses to contradictions for the neutral control targets, such as Hanks.

Now the crucial evidence that biased reasoning really is just that – emotionally biased rather than cognitively driven: The fMRI scans showed that the part of the brain associated with reasoning, the dorsolateral prefrontal cortex (DLPFC), was *not* involved when assessing these statements. Instead, the most active regions of the brain were those involved in processing *emotions* (ventro-medial prefrontal cortex-VMPFC), *conflict resolution* (anterior cingulate cortex) and *moral judgments* (posterior cingulate cortex). Furthermore, when the men reached completely biased conclusions (by finding ways to ignore information that contradicted their prior beliefs), brain circuits that mediate negative emotions (like sadness and disgust) diminished in activity whereas circuits that mediate reward became active. The lead scientist on this study summarized their results this way:

> We did not see any increased activation of the parts of the brain normally engaged during reasoning. What we saw instead was a network of emotion circuits lighting up, including circuits hypothesized to be involved in regulating emotion, and circuits known to be involved in resolving conflicts. None of the circuits involved in conscious reasoning were particularly engaged. Essentially, it appears as if partisans twirl the cognitive kaleidoscope until they get the conclusions they want, and then they get massively reinforced for it, with the elimination of negative emotional states and activation of positive ones.

Results like these bring home the point that we are prone to confirmation bias when seeking the truth through hypothesis testing. In order to overcome this inherent bias, we need a method of inquiry that counterbalances this tendency in much the way counterweights are used to balance a heavy load. The solution to this dilemma is called the hypothetico-deductive method of scientific inquiry, and it is the outcome of some of the most beautiful and clear thinking on the part of philosophers and scientists over the course of centuries. It begins with our facing the fact that when it comes to testing our beliefs, we are inherently biased, and then it asks what we can do about it. It ends with a scientific methodology that involves testing a hypothesis – a belief – by performing repeatable experiments that are designed to seek evidence *contradicting* that hypothesis. It is well worth examining the history of this struggle to come to terms with inherent observer bias.

Science: How We Got Here

Inquiring minds want to know the truth, and to borrow a phrase from Charles Peirce (1877), "inquiry is a means of fixing belief." Inquiry begins with a state of uncertainty and moves toward a state of certainty. We terminate the inquiry when we are satisfied that we know what we set out to know – the truth.

There are many ways of reducing the uncertainty of our beliefs. We could, for example, follow Plato's (429–347 BC) lead and engage in reflection and reason, the only means by which he believed the true nature of things is knowable. We could instead, like Plato's most illustrious student, Aristotle (384–322 BC), champion the idea that truth can be discovered through experience, particularly through careful observation and measurement. We could go even further, like Muslim scientists such as Alhacen (965–1040 AD), and engage in the practice of controlled experimentation to pursue the truth. If we choose to follow the lead of Aristotle and Alhacen, we follow the path of *scientific inquiry*.

The heart of scientific inquiry is *hypothesis testing*. In hypothesis testing, we scrutinize our beliefs by using them to make predictions about events that are directly observable. These observations constitute *evidence* that our hypotheses (beliefs) are either true or false. Sir Francis Bacon (1561–1626) argued that what distinguished science from other means of inquiry was exactly this reliance of verification.

According to Bacon, one "proved" the truth of a particular hypothesis by accumulating observations *consistent* with the hypothesis. Repeated confirmations lead to "laws" – descriptions of the rules that "govern" the natural world. He also introduced the notion of formulating *alternative* hypotheses and conducting explicit tests to distinguish among them. He believed this method yielded the greatest progress in scientific understanding. Bacon's method, therefore, rested primarily on the accumulation of confirming evidence for one's hypotheses.

Given the studies described in the introduction to this chapter, you may be skeptical of his position. You are not alone. David Hume (1711–1776) also questioned whether such confirmation was possible. He argued instead that no volume of "consistent observations" could be confidently regarded as "proof" of the truth of an hypothesis. As support he proposed the now-famous "black swans" argument: No amount of inductive evidence could prove the proposition *all swans are white* true, because somewhere, there could be an as-yet-undiscovered population of black swans. The mountain of evidence shored up to prove this hypothesis merely proved that it was supported *so far*. (As it turns out, there are black swans – in Australia!)

Karl Popper (1902–1994) took this objection even further, dismantling confirmation as the cornerstone of hypothesis testing. Instead, Popper emphasized "disproof" over proof and falsification over verification. He argued that truth is discovered by finding out what is not truth. According to Popper, no amount of evidence can "prove" the absolute truth of an assertion, but the relative truth of that assertion (i.e., the confidence it engenders) increases as the instances of "failed falsification" begin to accumulate. Put another way, the more one fails to disprove a hypothesis, the more confidence one has that the hypothesis is true. Sometimes a single experiment is sufficient to disprove an hypothesis. A single black swan falsifies our belief that all swans are white. Or as Einstein once put it, "No amount of experimentation can ever prove me right; a single experiment can prove me wrong."

Prove I'm Wrong, or Give Me the Most Bang for the Buck?

If falsification is of crucial importance, why do people generally approach hypothesis testing with a bias for employing a strategy that seems more consistent with discovering confirming evidence? Perhaps the answer is because it is an extremely efficient hypothesis-testing heuristic, even though, as we saw in the 2–4-6 task, it can lead to systematic errors.

The prevalence and implications of this type of strategy was beautifully described in a paper by Klayman and Ha (1987). Consider Wason's 2–4-6 task again. Wason had a rule in mind that neatly described a set of numbers, which we can call the target set (TS). His rule "any ascending series" was the target rule (TR) that exactly described that set. It was the rule you were supposed to figure out by generating tests. You constructed your best guess about the target rule based on the first triple you were shown. Your best guess is the hypothesized rule (HR). Your hypothesized rule exactly described a set of numbers as well. If your hypothesized rule was "ascending even numbers," then number strings that satisfied your rule constituted your hypothesized set (HS).

Now your goal is straightforward: You want to bring your hypothesized rule in line with the correct rule (HR = TR). That would mean that your hypothesized set would exactly match the target set (TS = HS). You could then perform one of two types of hypothesis tests. You could propose a triple that you believe had the properties of the target set (e.g., 6, 8, 10). That is called a positive hypothesis-test strategy. Or you could propose a triple that you thought did not have the properties of the target set (e.g., 2, 4, 7). That is called a negative hypothesis-test strategy. Wason found that people were far and away more likely to choose a positive test strategy. His subjects chose to test hypotheses using tests that they presumed possessed the properties of the target set (e.g., even numbers).

Klayman and Ha argued that under certain conditions – conditions that are actually very realistic – adopting a positive test strategy will allow you to *choose experiments that will provide the most conclusive evidence*. Another way to put this is that those experiments will, under certain conditions, allow you *to maximize expected information gain*. Hypothesis testing requires effort, so you are very clever if you prefer experiments that are expected to give you the most information for the least effort.

The most important of these special conditions is the *rarity condition* – that is, when *what you are investigating is a rare phenomenon*.

When this is the case, adopting a falsification strategy will be a little like trying to find a needle in a haystack. That is because the set of items that possess the properties that match the correct rule is small, but the set of items that do not possess those properties is extremely large. Under conditions of rarity, it is far more productive to examine items that possess your hypothesized properties.

Consider the case of depression. Imagine that you are the first person to suspect that one of the causes of depression is low serotonin levels. Your hypothesized rule would be something like "If serotonin levels are low, then depression results." Imagine that you are given the opportunity to examine four patients who come to a medical clinic in order to test your hypothesis.

The first patient has low serotonin levels, the second has normal serotonin levels, the third has been diagnosed with depression, and the fourth is not depressed but has come to the clinic for other reasons. You don't have much time, so you must choose which patient or patients to examine. Which would you choose?

This version of hypothesis testing is called the Wason Card Selection Task because it was first proposed and investigated by Peter Wason (1968) – the same researcher who gave us the 2–4-6 task. If you are like most people in Wason's study (and in countless studies since then), you chose to examine the patient with low serotonin levels to see whether the patient was depressed and the patient who was depressed in order to discover whether the patient's serotonin levels were low. Notice that you may be accused of engaging in confirmation bias. After all, you have selected just those cases that could confirm your hypothesis. Suppose you examine the depressed patient and find that their serotonin levels are low. This would be consistent with your hypothesis. But what if the patient's serotonin levels were normal? Does this disprove your hypothesis? Not really. Your hypothesis doesn't state that low serotonin is the only cause of depression. The patient may be depressed for other reasons.

But what if you had chosen to examine the patient who was not depressed, and it turned out that patient had low serotonin levels? You would have gotten incontrovertible proof that your hypothesis is false.

So this seems to be a more efficient test – one that is expected to be more informative.

In fact, a logician might argue that your hypothesis is a truth-functional conditional of the form P->Q, and the truth function for a conditional statement is false only when the antecedent (P) is true but the consequent (Q) is false. Therefore, an optimal hypothesis-testing strategy would be to examine the patient with low serotonin levels to test for depression and to examine the serotonin levels of the patient who is NOT depressed. Finding low serotonin levels and no depression would provide incontrovertible proof that the conditional was false. But, as you saw, that is not what most people do. In fact, this strategy probably didn't even occur to you. Is this rational?

Let's look at this issue from the perspective of employing a positive test strategy. This strategy makes sense if the phenomenon you are interested in is rare. What if the cases represented not individual patients but experiments that you could conduct? Adopting a falsification strategy would mean checking the serotonin levels of people who were not depressed to try to falsify your hypothesized rule. But, relatively speaking, depression is rare. The depression rate for American adults aged 18 and older is about 10%. Thus 90% of Americans are not depressed. If you took a random sample of 1,000 people, you would expect that only about 100 of them would be depressed and 900 would not. It would be more efficient to check the serotonin levels of depressed people because the set is smaller, and it contains people who have the property of interest: depression. If you found that serotonin levels in this group seemed unusually low, this would implicate low serotonin levels in depression. Your work, of course, would not end there. You would want to conduct experiments (as we'll see below). But at this stage of game, looking at the small positive test set is more likely to provide useful information with less effort than looking at the very, very large negative test set.

According to Oaksford and Chater (1999), you were not necessarily irrational when you focused on seeking potentially confirmatory evidence. Instead, you were demonstrating a bias for using a positive test strategy when testing hypotheses *so that you could maximize expected* *information gain.* This is consistent with the theory of *optimal data selection* (ODS) from Bayesian decision making (see Chapter One), rather than on Popper's falsificationism.

Oaksford and Chater (1999) propose six steps for testing hypotheses in their Optimal Data Selection (ODS) model:

(1) Goals: People adopt as a goal selecting the data that has the greatest expected informativeness (*EIg*) about whether the rule is true (or whether the antecedent (*p*) and consequent (*q*) of the rule are independent).

(2) Environment: People typically engage in hypothesis testing when something out of the ordinary happens. This means that when they start up their hypothesis-testing machinery, they usually already assume that the properties of interest are rare in the environment (e.g., depression and low serotonin levels are rare).

(3) Computational limitations: If we had limitless time and money, we could do all possible investigations to get at the truth. But we usually don't. Instead, we have to make decisions in real time with limited resources.

(4) Optimization: People will choose to select the most informative data they can. Assuming rarity in the Wason task leads to the following order of expected information gain across the four cards: $EI_g(\mathrm{P}) > EI_g(Q) > EI_g(\mathrm{N}ot\text{-}Q) > EI_g(Not\text{-}\mathrm{P})$. That is why the necessity of investigating the P case (depression) and the Q case (low serotonin) seemed so obvious to you.

ODS also makes the novel prediction that selection-task performance should change if the rarity assumption is manipulated. This has been tested and observed. For example, if people are given depression cases and normal cases to investigate, the probability that they will select the normal cases increases as the number of depressed cases increases. In other words, people choose to investigate rare events, and if the normal case seems rare, they will choose to investigate those cases rather than the increasingly more numerous depression cases (Oaksford and colleagues, 1997).

Ok, Show Me I'm Wrong

Let's return to your hypothesis concerning serotonin levels and depression. Suppose that you have examined a number of depressed individuals (rare cases) and have noticed that on the whole, they tend to have low serotonin levels. Your hypothesis seems to be plausible based on this information. But you don't want to stop there. You want more proof.

After thinking about it, you realize that your hypothesis implies that increasing serotonin levels should actually alleviate depression. So you design a drug that interferes with the clearing or "re-uptake" of serotonin in the brain, and call it a serotonin re-uptake inhibitor (SSRI). Testing the efficacy of this drug will not only perhaps provide a treatment for depression, it will also test the truth of your belief that depression is actually caused by low serotonin levels.

Whether you realize it or not, you have just begun implementing the modern hypothetico-deductive method of scientific inquiry, which can be summarized as follows:

1. Identify a problem.
2. Form a conjecture (*hypothesis*) to explain the problem.
3. *Deduce* prediction(s) from your hypothesis.
5. Design experiment(s) to test each prediction.

If you proceed as stated, however, you run the risk of seeking confirming evidence all over again. So you tweak this a bit to allow yourself the opportunity to engage in falsification. You do this by constructing a null hypothesis that is *mutually exclusive* with your favored hypothesis:

H_0 : SSRI has NO effect on depression
H_A : SSRI DOES have an effect on depression

Notice that both these hypotheses can't be true at the same time – they are mutually exclusive. You set up a study by assigning depressed participants into two groups, those who take the SSRI and those who receive a placebo (or sugar pill). Then you assess levels of depression in the two groups. The placebo group is particularly important because the "placebo effect" is a real effect: People sometimes get better just because they think they have received a drug. In the case of some illnesses, such as ulcers, as many as 50% of placebo takers may notice a genuine improvement in their condition. So for you to safely conclude that your SSRI is really working, you must compare the improvement rates of the SSRI group to those of the placebo group.

Notice also that your favored hypothesis H_A is worded so that you will be looking for changes in depression in either direction – improving or worsening. This is called a *two-tailed test* because you are looking for changes in either direction. If there were no difference, you would not be able to reject the null hypothesis, and you would conclude that your SSRI was ineffective. But if the SSRI group's scores were lower than the placebo group, that would mean your SSRI lowered depression levels. You could then conclude that your SSRI actually lessened depression. On the other hand, if the SSRI group scores were higher than the placebo group's, it would mean your SSRI raised depression scores – it made depression worse. Stating your null hypothesis this way lets you look for differences in either direction via statistical tests.

If you had chosen to state your favored hypothesis as "My SSRI alleviates depression," then the mutually exclusive null hypothesis would be "My SSRI has no effect or makes depression worse." This would be *a one-tailed test* because you are looking only for improvements due to your SSRI. If you rejected this null hypothesis, it would tell you that your SSRI group had less depression. But if your SSRI actually made depression worse, you would miss that because your null hypothesis lumped that possible outcome together with "no effect." All you would be able to say is "the results show that my SSRI doesn't alleviate depression, but I can't say whether that is because it has no effect (SSRI group and placebo group are the same) or it actually made depression worse (SSRI group has higher depression scores than the placebo group).

So you now have measures of depression for each participant in each group. Because people are not carbon copies of one another, the measures even within the groups can differ from one another. If we were to take the (Fahrenheit) body temperatures of a million people, we would find that they all differ from one another slightly. Some have a body temperature of 98.4 degrees, some 98.9 degrees, and so on. If we randomly divided these people into groups and compared the average temperatures of each group, we would find that the averages differed from one another as well. One group might have an average temperature of 98.4 degrees, another 98.8 degrees, and so on. But if we took the average of all of those group averages, we would find that the most frequent temperature was 98.6 degrees, the true average human body temperature. In other words, the group averages would distribute themselves around the true population average. This is called the *sampling distribution of the mean* (or average). If our samples were truly random, you would also find that the larger size of the sample group, the closer its average is to the true population average. The average score of smaller sample groups are more likely to bounce all over the place.

The sampling distributions of many traits (such as temperature) look like a bell-shaped curve, which allows us to do something mathematically interesting and useful. We can use these distributions to limit the probability that we will make an incorrect decision. The "bell" shape occurs because of the way the scores distribute themselves around the average, as shown in Figure 7.1.

Think of a standard deviation (SD) as the typical distance that separates a score from the average. About 34% of the scores will fall between the average and one standard deviation above the average, and about 34% of the scores will fall between the average and one standard deviation below the average. Another 14% will fall between one and two standard deviations above the average, and the same will be true below the average. Almost 2% will fall between two and three standard deviations above average. Ditto for below. So all told, about 95% of scores will fall between +2 and −2 standard deviations from average.

Let's see how this works to test the hypothesis that someone has a fever. We'll use .6 as a standard deviation, which is about right. We have three people, A, B, and C, and their temperatures are shown on the bell curve in Figure 7.1.

A's temperature of 99 degrees is not that far from 98.6 degrees and falls well within the range of normal temperatures. The null hypothesis of "normal temperature" seems warranted. B's temperature of 100.1 degrees,

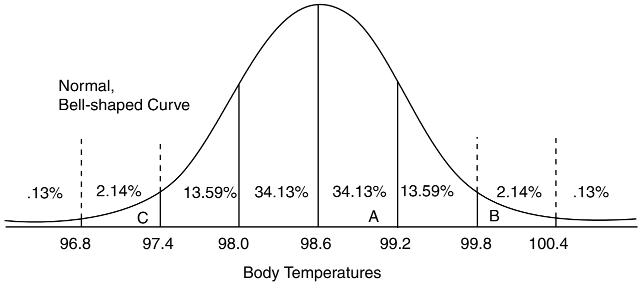


FIGURE 7.1. The "Bell" curve for normally distributed data.

on the other hand, falls outside of the normal range of temperatures. Only about 2.5% of healthy people have a body temperature that high naturally, so the null hypothesis of "normal temperature" does not seem appropriate. The same can be said for C's temperature; only about 2.5% of the healthy population has a temperature below 97.4 degrees naturally, so the null hypothesis of "normal temperature" does not seem appropriate. You would conclude that B and C are probably sick.

But here's the problem: You *could* be wrong. If B and C really just happened to fall in that 2.5% of the population that naturally have high or low body temperatures, that would mean you *rejected a true null hypothesis*. This is called a *Type I Error* or a *false positive*. The probability that you would make this kind of mistake is about 2.5%, because 2.5% of normal people have temperatures that high or higher. So your chances of having made an incorrect decision would be about 2.5 chances out of 100. This is referred to as alpha (α). Alpha also can be considered an estimate of your *uncertainty* in the repeatability of your results.

You could also make another type of error, namely, failing to reject a false null hypothesis. Using our example, that would mean that you decide the person with a temperature of 101 degrees just normally has a temperature that high and isn't running a fever. But suppose you're wrong – the person really is sick. That means you *failed to reject a false null hypothesis*. This type of error is called a *Type II error* or a *false negative*. The probability that we will commit this kind of error is referred to as beta (β). Table 7.1 lays it out clearly.

The point is that, when doing scientific investigations, you can never know the truth with certainty. You are always in a position of

Table 7.1. *The Logic of Hypothesis Testing*

Decision	Truth of the Matter	
	H_0 True	H_0 False
REJECT H_0 Decide patient has normal temperature	TYPE 1 ERROR!	CORRECT DECISION
FAIL TO REJECT H_0 Decide patient has abnormal temperature	CORRECT DECISION	TYPE II ERROR!

making your best guess based on the information available. To compensate, it is best to try to maximize certainty by minimizing the probability that you will make an error. Rejecting a true null hypothesis is usually a serious error, and the convention is to set alpha very low at 5%. If your null hypothesis was set up as a one-tailed test, you would apply the whole 5% to the extreme high end of the distribution. If your null hypothesis was set up as a two-tailed test, you would be interested in two places on the distribution of scores, one that marks off the 2.5% of the scores that fall at the extreme high end of the distribution and the 2.5% that fall at the extreme low end. These are called the *critical regions for rejecting a null hypothesis.*

Knowing all this, you decide to apply the same logic when testing depression scores in your groups (assuming that depression scores distribute themselves in a bell-shaped curve). The average depression score in your placebo group can serve as the "normal" or "expected score," just as 98.6 degrees constitutes a "normal" or "expected" body temperature. You compare the average depression score in the treatment group with the average depression score in the placebo group. You use the size of your groups and the variability of the scores within each group to calculate the probability that the difference you observed between the groups was normal variability. You can reject the null hypothesis ("no difference") only if your results fall within the rejection regions. If they fall in the upper rejection region, you can conclude that your SSRI increases depression. If they fall in the lower rejection region, you can conclude that your SSRI lowers depression.

Suppose you compare the groups and decide that they look pretty different. In fact, the difference is large enough that they fall inside one of your rejection regions. But suppose the truth of the matter is that your SSRI is NOT working. The difference is due to chance factors or measurement error. Then you have made a Type I error – you incorrectly rejected a true null hypothesis. Your SSRI drug doesn't really work, but you conclude that it does.

Suppose instead that we compare the groups and decide they look pretty much the same. There is a difference, but it is so small that our results do not fall in your extreme critical rejection regions. You decide the SSRI is not working. But suppose the truth is that it *is* working. It produces a small but true difference between the groups, and perhaps larger doses or a longer time frame would make a genuinely significant difference. You have committed a Type II error by *failing to reject a false null hypothesis*.

Let's suppose that this is what you actually observe. Your results don't fall in the rejection regions, so you cannot reject the null hypothesis statistically. Do you now conclude that your SSRI is ineffective? Is this what scientists really do?

As it turns out, we don't need to speculate. This example is based on a real-life case. A team of researchers led by University of Hull's Irvin Kirsch and colleagues (2008) obtained all clinical trials submitted to the U.S. Food and Drug Administration for the licensing of four SSRIs (Prozac, Effexor, Paxil, and Serzone). They found that these drugs helped severely depressed people but were not better than placebos for mild or moderate depression. The researchers concluded that "although patients get better when they take antidepressants, they also get better when they take a placebo, and the difference in improvement is not very great. This means that depressed people can improve without chemical treatments." But Mary Ann Rhyne, a spokeswoman for the maker of Paxil, objected to this conclusion because it was based only on clinical trials submitted to the FDA. According to Rhyne, "The authors have failed to acknowledge the very positive benefits these treatments have provided to patients and their families who are dealing with depression and they are at odds with what has been seen in actual clinical practice." (Quotes obtained from Reuters News, Tuesday, February 26, 2008, 6:24 AM ET). Notice that Rhyne's rebuttal rests on pitting what doctors "see" and what the scientific data show.

Box 7.2. Statistics or Bayes?

As we saw in Chapter Three, there is frequently a disconnect between scientific studies and clinical decisions. A physician wants to know what is the probability that my patient has a particular disease? A scientist wants to know whether a particular factor is a cause of a disease. The physician must make a Bayesian decision; the scientist must do an experiment and test for statistical significance to reject null hypotheses.

Here is a beautiful example of how these different types of approaches lead to different answers even in the sciences (taken from Anderson, 1998):

Suppose that an ecologist is looking for Marbled Murrelets in a forest. What is a good sign that Murrelets might be present? A possible candidate is the existence of suitable nest sites: big, horizontal branches with a good layer of moss and lichen. The investigator designs a quick, standardized survey to detect those branches and has tried it out on 1,000 stands, producing the data shown here:

	H_0 true: Murrelets are NOT nesting in the stand	H_0 false: Murrelets are nesting in the stand	Total
Survey data: Potential nest sites were NOT seen	808	4	812
Survey date: Potential nest sites were seen	142	46	188
Total	950	50	1000

H_1: Suitable nest sites are associated with presence of nesting Murrelets
H_0: Suitable nest sites are NOT associated with presence of nesting Murrelets

$$\chi 2\ (1) = (50{*}812)/1000 = 41$$
$$(50{*}188)/1000 = 9$$
$$(950{*}812)/1000 = 771$$
$$(950{*}188)/1000 = 179$$
$$= (4-41)^2/41 + (46-9)^2/9 + (808-771)^2/771)$$
$$+ (141-179)^2/179 = 186$$

The probability of making Type I error if you reject H_0 less than .0001! So suitable nest sites certainly are good places to look for Murrelets.

(*continued*)

But what is the actual probability that Murrelets are nesting given the fact that you've found potential nest sites? That's easy: 46/188 = .24. So are stands a good indicator of nesting? No; your chances of finding Murrelets in the stands are a little less than 1 out of 4 (25%). In a medical context, this means that Bayesian reasoning is more appropriate for asking questions like "How likely is it that this patient has this disease?" and statistical significance tests are more appropriate for asking "Is the factor a likely cause of this disease?"

Stopgaps and Backup Systems

As this chapter shows, hypothesis testing is a thorny process fraught with biases that may or may not be irrational. Does this mean that we should abandon it as a means of inquiry?

To reduce the chance of error (or fraud) in hypothesis testing, there are several "backup systems" in place in the scientific community. First, it is common practice for other scientists to attempt to repeat experiments in order to duplicate the results. In the example given here, independent laboratories would replicate the results using the same or modified study procedures. Replication is a crucial procedure for ferreting out scientific errors. Second, the methods and procedures undertaken in studies are typically subjected to peer review. Scientists submit written descriptions of their work to scientific journals. The editors of the journals then send the manuscripts to fellow scientists familiar with the field (called referees) for evaluation. The referees may recommend publication with or without suggested modifications or reject them on the basis of insufficient quality or importance. In some scientific journals, the rejection rate can be as high as 95%.

Does this system always get it right? Probably not. Scientists are human. Frequently, peer review blocks publication of poor quality work, but sometimes those pieces get published due to cronyism, favoritism, or potential for financial gain. Sometimes, good studies whose results can't be explained by current theories or studied via accepted means lead to genuine upheavals in the way scientists conceptualize a phenomenon or the way the phenomenon is studied. In his 1962 book *The Structure of Scientific Revolutions,* philosopher Thomas Kuhn (1922–1996) argued that science doesn't progress by way of a linear accumulation of new knowledge. Instead, he argued, it undergoes periodic revolutions where one way of thinking about a phenomenon is replaced with an entirely different way of thinking about it. He called these upheavals "paradigm shifts." They completely transform the nature of scientific explanations within a particular field.

In his scholarly study of how scientists go about their business, Kuhn noticed that scientific investigation takes place in three distinct stages. He referred to the first as pre-science,; a certain phenomenon captures the attention of a scientist or group of scientists, and they begin to tinker with it in order to understand and explain it. This stage lacks a central paradigm in that nobody yet knows how to conceptualize the phenomenon. This is followed by "normal science." After conducting enough experiments, scientists think they understand what is going on, and they propose a theory to explain it. This then becomes the central paradigm, the way everyone agrees the phenomenon should be explained and studied. But then a disconfirming result is reported. This result typically will not be seen as refuting the paradigm but as the mistake of the researcher. As disconfirming results start to accumulate, however, science reaches a crisis point. At this point, someone usually proposes an entirely different theoretical framework that explains the phenomenon in a different way and accommodates the previously problematic results. The new paradigm overthrows the old as more scientists adopt it. Kuhn called this "revolutionary science."

The important thing is that during the period of upheaval, there exist rival paradigms that are incommensurable; they explain the phenomenon in different ways, often making use of concepts that fit only one theoretical view. Kuhn also noticed that paradigms are overthrown when five criteria are met: (a) the new paradigm is more accurate than the old in its predictions, (b) it is internally coherent as well as consistent with other theories, (c) it is broader in scope than the old paradigm because it explains a wider variety of effects, (d) it is simpler than the old paradigm, and (e) it is fruitful because it makes new predictions that the old paradigm did not.

The history of science is full of these kinds of paradigm shifts. For millennia, scientists accepted Aristotle's and Ptolemy's paradigm of celestial mechanics – namely, that the sun and other planets revolved around the earth. Then in 1543, Nicholas Copernicus published his sun-centered theory. It threatened not just the Aristotelian/Ptolemaic view, but the Biblical view of the position and status of the earth as the center of the universe. The Vatican weighed in, putting Copernicus' book on its forbidden book list. It remained ignored for eighty years until Italian scientist Galileo found convincing evidence of the Copernican view using his own newly invented telescope. He saw that Venus had phases just like the moon (full, half, crescent, new). This could happen only if Venus were orbiting the sun, not the earth. He also saw the moons of Jupiter orbiting Jupiter – not earth. The Vatican again intervened and demanded that Galileo retract his finding or risk excommunication and possibly execution. Galileo did so, and then, frustrated that he was required to suppress this important truth, he published a dialogue in which he defended and supported the Copernican theory. The Vatican was swift in its response, demanding under threat of torture that he recant this absurd view. When he refused, he was put under house arrest for the rest of his life. But some ideas just won't die. Other scientists (notably German scientist Johannes Kepler, Dutch scientist Tyco Brahe, and English scientist Sir Isaac Newton) continued their celestial investigations, ultimately vindicating the Copernican view.

More recent revolutions include a shift from classical (Newtonian) mechanics to relativity and quantum physics, the replacement of miasma with the germ theory of disease, and the replacement of pangenesis with Mendalian inheritance. Psychology has undergone several paradigm shifts in its young history as we came to understand more about information processing through the invention of computing systems. In the late 1800s, structuralists thought the role of psychology was introspecting the quality and content of mental events such as sensations and thoughts. This idea was vehemently overthrown in the early 1900s by behaviorists who believed strongly that a real science must be grounded in observables such as behavior. At about the same time, computer scientists began creating machines that carried out intelligent functions, an enterprise that reached full bloom in the 1960s. The only way to describe what these intelligent programs were doing was to make reference to their internal states, such as rules, knowledge (data structures), and goal stacks. Asking themselves "If machines can have internal states, why can't we?", psychologists wrested control of the discipline back from the behaviorists, and the cognitive revolution was born – and continues to thrive today.

So how can we summarize the scientific process? Despite paradigm shifts and revolutions, the workaday fundamental core component of scientific investigation is the *hypothetico-deductive* method. Scientists pose hypotheses deduced from theories and test them using Popperian falsification. This method is widely accepted in the scientific community because it's the best system we've got. This means

- Good scientific questions can be tested empirically.
- Good tests are replicable and have the highest expected information gain.
- Scientific disciplines comprise sets of explanations or theories based on *broad consensus* concerning the meaning of replicated, empirical observations.

Because current scientific theories often rest on those very worrisome italicized words – *broad consensus* – some scholars have concluded that scientific theories are no more than cultural conventions rather than true descriptions of reality (see Spiro, 1996 for a review of post-modern analyses). This strikes many scientists as akin to concluding that there is no such thing as an elephant because the proverbial blind men investigating it all came back with different descriptions of it ("an elephant is like a rope" from the blind man who felt its trunk, "an elephant is like a tree" from the blind man who felt its leg, etc.). Scientific hypothesis testing is one of the strongest methods available for discovering truth. Daniel Bernoulli (1700–1782), for example, used scientific hypothesis testing to develop and articulate the principle that explains the physics of flight. To paraphrase Bertie, Alexander McCall Smith's (2007) precocious six-year-old protagonist, if you doubt the veracity of scientific reasoning, you might want to think twice about getting on an airplane.

EIGHT

Problem Solving

TURNING WHAT YOU DON'T WANT INTO WHAT YOU WANT

"*Houston, we have a problem.*" It's 1970, and NASA has just launched another manned space expedition, *Apollo 13*. It is the seventh manned expedition and the third scheduled to land on the moon. People have become fairly jaded about these expeditions, and so little air time is devoted to televising the mission – until an oxygen tank explodes, crippling the spacecraft and stranding its three astronauts 205,000 miles above the earth. While the astronauts search for ways to survive in the limited time they have left, NASA's heroic ground crew searches for ways to bring the craft safely back home using nothing more than the items available to the men on board. As we all know from history (and a really great Ron Howard movie starring Tom Hanks), they find a solution, and all ends well.

Although more dramatic than the usual problems we encounter in everyday life, the *Apollo 13* story contains all of the elements that comprise all problem solving, from solutions to the little annoying problems (finding your lost keys) to potentially life-changing ones (finding a cure for a deadly disease). The key insight derived from *Apollo 13* – and from a century of research on problem solving – is this: *Problem solving is search.* When you rummage around your house looking for your keys, you're involved in search. When investigators sift through a crime scene, they are searching for evidence to solve a crime. When a prosecutor questions a witness, she is searching the witness's memory for information relevant to solving a crime. When a medical researcher conducts an experiment on cancer cells, she is searching for mechanisms underlying the disease. When you reason through a problem, you are searching your knowledge base to find facts and inferences that lead to a solution. And when you wake up in the night with the solution to a problem that was bugging you all day, your brain was engaged in an implicit search process that led to the solution.

This seemingly simple insight that problem solving is search turned out to have enormous implications, not just for improving human problem-solving performance but also for automating it. Today, we routinely rely on automated systems to solve problems for us. Automated systems retrieve information for us on the Internet, drive robotic arms that assist in surgeries or build cars, and even make stock trades. And that is because we've learned a lot about how to describe and implement search processes. And we've learned a lot about how to characterize problems.

This is how Gestalt psychologist Karl Duncker (1903–1940) described problems and problem-solving in a famous monograph (published in 1945):

> A problem arises when a living creature has a goal but does not know how this goal is to be reached. Whenever one cannot go from the given situation to the desired situation simply by action, then there has to be recourse to thinking. (By action we here understand the performance of obvious operations.) Such thinking has the task of devising some action which may mediate between the existing and the desired situations. (p. 1)

It turns out the year 1945 was a banner year for the publication of problem-solving insights. This same year saw the publication of mathematician George Polya's (1887–1985) quintessential text for solving problems, the aptly titled *How to Solve It.* Here is how this brilliant mathematician summarized the problem-solving process.

1. First, make sure you understand the problem. You do this by developing a representation of the essential aspects of the problem. You do that by searching your knowledge base for information that seems to you to be solution-relevant.
2. After understanding, then make a plan for solving the problem. This will also usually involve searching one's knowledge base for solutions that are appropriate for the problem as represented.

3. Carry out the plan by executing your solutions.
4. Look back on your work and ask "how could it be better?"

As Duncker makes clear, we have a problem when our current state doesn't match our goal state. If we have no immediate action at our disposal to reach our goal, we typically resort to *thinking*. As Duncker and Polya both make clear, the goal of that thinking process is *searching* out means (or actions) we can put into play to bring us closer to our goal. So the process of problem-solving can be described as *difference reduction – employing actions that will reduce the difference between our current state and goal state.*

Imagine you want to go to a movie, but you can't find your car keys. Your goal state is being at the movie theater. Your current state is being at home. So you've got a problem: You are in a situation where *your goal state differs from your current state.* At this point, you will try to take actions that will get you to the movie theater. You might decide to call a friend to pick you up or walk there if the weather is nice and the theater isn't too far away. Using problem-solving parlance, these are *means or operators* that you can implement *to reduce the differences between your goal state and your current state.* You know you no longer have a problem when your current state (you are at the movie theater) is the same as your goal state (you are at the movie theater). Problem solved.

And if the problem is one that can't be solved by searching the environment? That's when we resort to thinking and imagination.

When Problems Are Well-Defined

Some problems are *well-defined* – that is, *they have a clearly specified start and goal state and clearly defined solution paths.* It's pretty easy to get to a theater once you have a clear set of directions or to figure out what twelve divided by four is once you know the rules of division or how to make scrambled eggs if you have a recipe.

The nice thing about well-defined problems is that there usually exist algorithms for solving them. *An algorithm is a procedure or formula that terminates in a result,* such as trying to find your keys by searching systematically through every space in every room in your house. You will get to the last room sooner or later (the procedure will end in a result), and if the keys you're looking for are in the house, you are guaranteed to find them. If they aren't, the procedure will still end because you'll come to the last room. Most algorithms for well-defined problems not only guarantee termination with a result, they guarantee the right result. If you follow the rules of multiplication accurately, you will absolutely obtain the correct answer to the problem "296 times 398" in finite time. If you follow a recipe for "scrambled eggs" and do every step as described, you will end up with scrambled eggs in finite time.

Let's be more specific about this problem-solving technique: *An algorithm is a series of specific steps that solve a particular problem.* We can define each term in that statement in the following way:

- **Step**: Each action you must take is called a step. If you're following a recipe for scrambled eggs, one step would be cracking the eggs.
- **Series**: The steps must be done in a particular order, and each of the steps must be used (unless the algorithm says otherwise). So you must crack the eggs before attempting to scramble them.
- **Specific**: A step must NOT be replaced by a similar step. So you must crack the eggs, not crush them.
- **Solve** – An algorithm produces a final result, which is the solution to a problem. If you follow the recipe exactly, you will end up with cooked scrambled eggs.
- **Particular problem**: The algorithm for one problem will not usually solve a different problem. This recipe is specific for scrambled eggs. If you want to bake a cake, you need to find another algorithm (recipe).

A particularly powerful and efficient class of algorithms is recursive. Procedures are **recursive** if they satisfy two conditions: There must be a base case, and the rules systematically reduce all other cases toward the base case. A simpler way of putting this is *a recursive procedure employs rules that make reference to themselves*. For this to work, the following conditions must also apply: The problem must be solvable, and there must be a terminating clause (you have to know when you're done).

A classic example of a problem that can be solved with a recursive algorithm is the Tower of Hanoi (see Box 8.1). You start this game by stacking one or more discs on the farthest left peg in order of increasing size, with the largest on the bottom. You then have to transfer all of the discs to the farthest right peg, organized in the same way – in order of increasing size with the largest on the bottom. The catch is that you can transfer only one disc at a time, and you can never put a larger disc on top of a smaller disc. This seems easy, but take a bit of time now to try it. It will drive you crazy!

Box 8.1. The Tower of Hanoi

The Tower of Hanoi is a type of problem that can be solved using a recursive procedure. As you can see from the figures below, there are three pegs. Your job is to transfer the discs from A to C, one at a time, without ever putting a large disc on top of a smaller one.

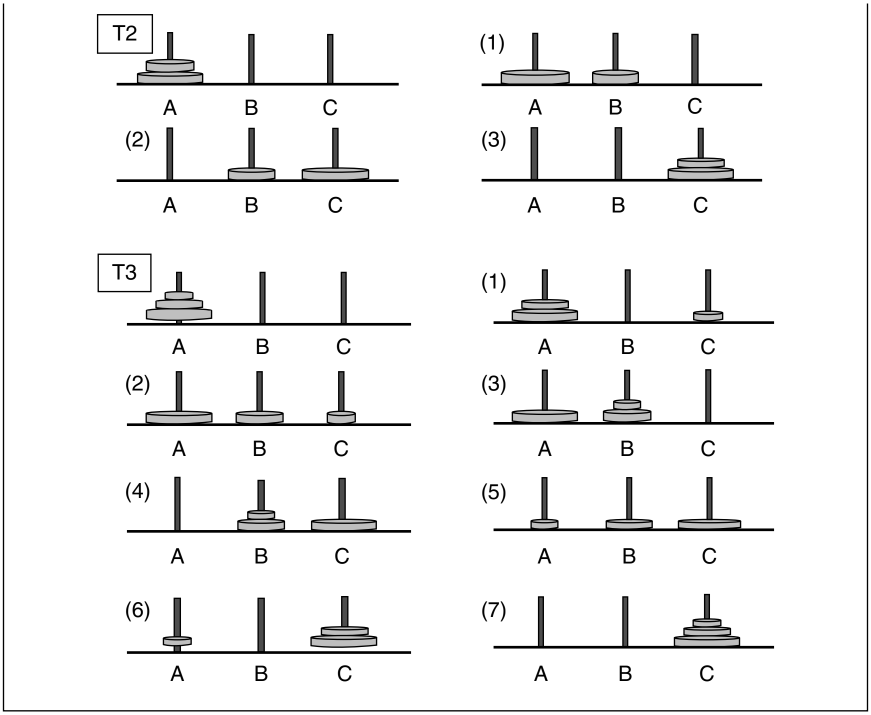
The recursive algorithm for solving the problem can be summarized like this:

1) Move the top N-1 discs from A to B.
2) Move the Nth disc from A to C.
3) Move the N-1 discs from B to C.

The trick is to see that every time you move a disc, you have a new problem with a smaller number of discs than can be solved with the same procedure. So once you master this algorithm, you can solve it with any number of discs.

If there is only one disc, it only takes one move: You simply transfer it from A to C. If there are two, then it takes three moves. Move the small one from A to B, the large one from A to C, and then the small one from B to C – just as the algorithm says. This is illustrated in T2 below. Three discs require seven moves, as shown in T3 and as described by the algorithm.

Now before you get cocky and enter a tournament with, say, 10 discs, keep this in mind: The formula for figuring out how many moves it will take is simple: $2^n - 1$, where n is the number of discs to be moved. For example, when n=3, then the formula is $2^3 - 1 = 8 - 1 = 7$. Ten discs would require $2^{10} - 1 = 1{,}024 - 1 = 1{,}023$ moves!



The recursive procedure for Tower of Hanoi is actually very simple: To move *n* discs from A to C using B as spare, do the following:

- if *n* is 1 (there is only one disc), move the disc from A to C
- otherwise
  - o move *n*–1 discs from A to B, using C as spare
  - o move one disc from A to C
  - o move *n*–1 discs from B to C, using A as spare

Now all you have to do is exactly the same thing for *n*–1 discs, and so on, until there is only one disc remaining and you can execute the first step. So if there were three discs on peg A, you would execute the steps following "otherwise." Then you would have two discs on peg A and one on peg C, and you would go through the whole procedure again, leaving two discs on peg A and two on C. Since there is only one peg on A, you could execute the first step – and you're done! This is recursive because the procedure applies itself repeatedly until the rules reduce multiple disc cases to a single disc case and then ends with the final move of that disc. The solution is shown in box one.

Rubik's cube is another problem that is defined by sets of recursive procedures. Once you know what they are, this seemingly difficult game becomes trivially easy to solve.

If you're beginning to think that recursion pertains only to geeky games, you may be surprised to find that you engage in recursion every day when you speak or understand language. Natural languages typically consist of sets of recursive rules for generating or parsing sentences. In fact, this is one of the most powerful aspects of natural language; it makes it possible for you to generate an infinity of sentences using a few simple rules. Think about it: You've never read the sentences printed here before, yet you have little difficulty understanding them. And every day, you speak hundreds of unique sentences that you've never generated before. And you do this effortlessly, often while you are simultaneously doing something else.

To see how this happens, consider this very simple set of recursive "rewrite" rules for English. If you follow these simple rules, you can build grammatical sentences in English of any length. The items in brackets are optional; everything else is mandatory. The arrows (->) mean "is made up of."

sentence -> **noun phrase** + **verb phrase**
noun phrase -> [**article**] + [**adjective**] + **noun** + [relative clause]
verb phrase -> **verb** + [**adverb**] + [noun phrase]
relative clause -> **relative pronoun** + **verb** + noun phrase

Notice that the rules specify how to create larger units (sentences, noun phrases, verb phrases, and relative clauses) from smaller units (nouns, articles, adjectives, adverbs, pronouns) and that they refer to each other in their instructions. Using these rules, we can generate the following sentence:

The mystery thief who stole the car robbed a convenience store today.

"The mystery thief" is a noun phrase (article, adjective, noun); "robbed a store" is a verb phrase (verb, noun phrase); "who stole the car" is a relative

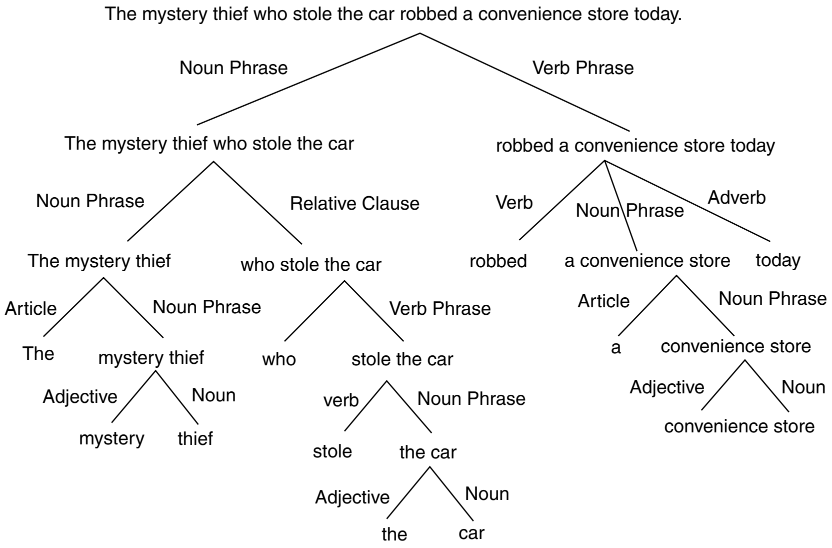


FIGURE 8.1. An English sentence parse tree for a complex sentence constructed from four simple rules.

clause (relative pronoun, verb, noun phrase). Figure 8.1 presents the parse tree for this sentence, including which rules generated which parts.

We could keep adding components according to the recursive rules and end up with sentences like this:

> *The mystery thief who stole the car that belonged to the French ambassador who was visiting his uncle robbed a convenience store today.*

When Problems Are Not So Well-Defined

If life consisted mostly of well-defined problems that came packaged with the algorithms to solve them, we would be very happy campers. Unfortunately, the problems that define our lives are often *ill-defined* – that is, *they do not have clear goal states or well-defined solution paths*. How do you make enough money to save for retirement, how do you avoid war, or how do you get that girl or guy to go out with you? You have an algorithm for making scrambled eggs, but how do you create a new egg-based dish that will please your breakfast guests? You have implicit mastery of the algorithms for generating English sentences, but what should you say to persuade someone to your point of view?

There is no agreed-upon solution path for reaching any of these goals. There may be many such paths or none at all.

Ill-defined problems don't lend themselves to clear solution procedures. For this reason, we frequently use heuristics to try to solve them. *Heuristics are "rules of thumb" or experience-based strategies for increasing the probability of solution success.* Heuristics are not guaranteed to end in a result, nor do they guarantee correct results. Sometimes we use heuristics even when an algorithm is available to solve a problem. What is the solution to the problem "96 times 58?" If you have a calculator handy, you could employ that means to achieve a solution. But what if you don't? You could laboriously plod through the multiplication algorithm to arrive at the solution. But what if you're trying to figure out whether you can afford to buy 96 yards of carpeting that costs $58 a yard? A heuristic would work just as well: Round 96 to 100. Now the problem is easy: Just add two zeroes to $58, and you've got your answer: $5,800. These simple heuristics take a mentally taxing problem and reduce it to a problem that is quickly solvable with very little mental effort.

As this example shows, heuristics usually have two features that make them enormously useful and practical: *They usually require less effort and less time to implement.* In the perfect world, we would have unlimited time and processing resources to find solutions. In the real world, we frequently don't. We have to find solutions in real time with reasonable effort. Hence, heuristics often are our friends.

We can revisit the earlier example of finding your keys in order to distinguish between algorithms and heuristics. Suppose you're getting ready to leave for work, and you can't find your keys. Now what? You will, of course, begin to search for them. If you approached this problem "algorithmically," you might systematically search your house, beginning in the attic and working your way through each room on each floor until you reached the last room in your basement. If you searched each room on each floor before taking the stairs down to the next floor, you would be doing a *breadth-first search*, as shown in Figure 8.2.

If there were staircases in some of the rooms that led directly to rooms directly below, and you searched each room along those staircases

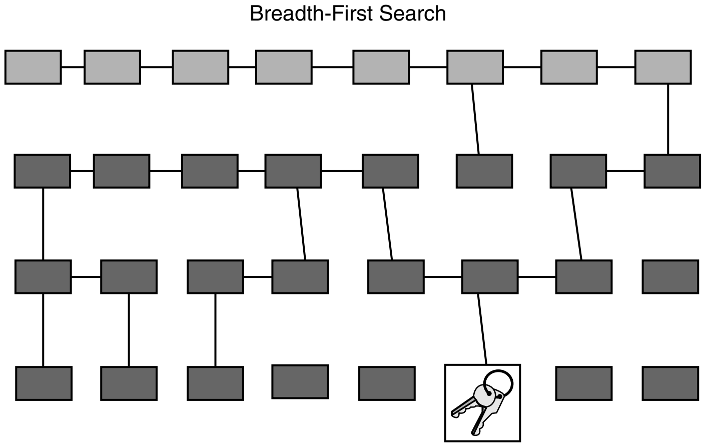


FIGURE 8.2. When doing a breadth-first search, all the possibilities at a given level are searched before going to a new level.

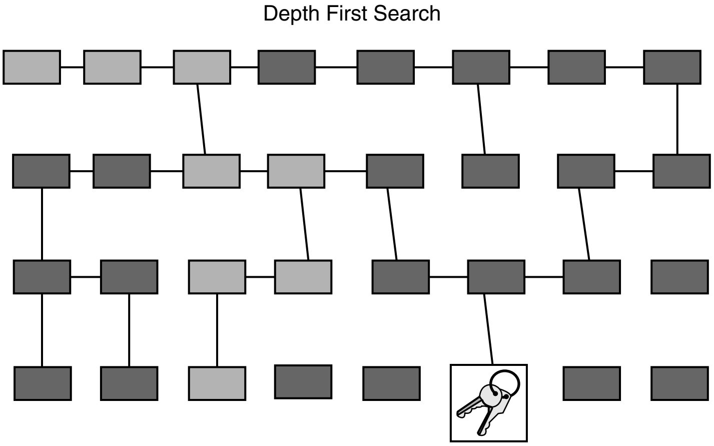


FIGURE 8.3. When doing a depth-first search, all the possibilities connected to lower levels are searched. Here, the searcher would have to go back up to the top level and start down a new path.

before returning to the attic to search remaining rooms, you would be doing a *depth-first search*, as shown in Figure 8.3.

Following either of these search procedures would guarantee that you would (a) come to the end of the search and (b) find your keys (if they were indeed in the house). But this would be very time consuming.

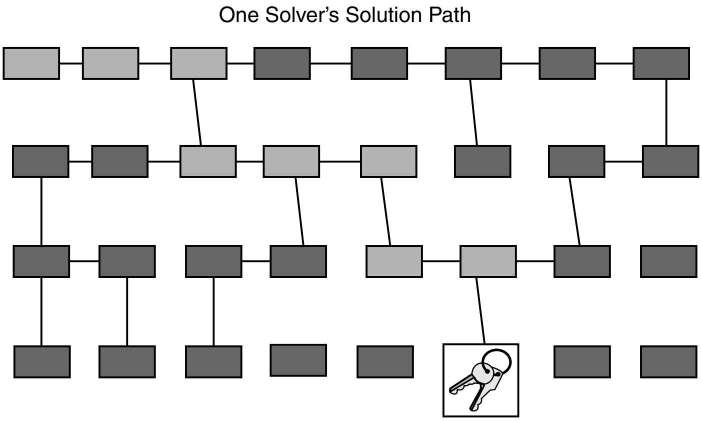


FIGURE 8.4. How one problem solver searched for a solution.

What if instead you used a heuristic? You would begin by looking in the last place you are certain you had them and re-trace your steps forward in time from there. You remember unlocking your door last evening and running to answer the ringing telephone. So you search the route from the front door to the telephone. Not there? Where did you go next? To the bedroom. Search there. And so on. Unlike systematically searching every room in the house, this heuristic does not guarantee that you will find the keys even if they are in the house. But it takes less time and less effort. Furthermore, it turns out to have a very high success rate. (Many thanks to Agathie Christie's Hercule Poirot for describing this heuristic.)

No matter what problem you're trying to solve, we can formally describe it as a search through problem space, where *"problem space" refers to the start state, goal state, and all possible intermediate states that pertain to that problem.* In our key-searching example, this means every possible room you can choose to search and the paths that lead from that room. For complex problems, the problem space can be so large as to be intractable. An actual problem solver's actual solution path is usually a subset of the problem space, as shown in Figure 8.4.

The best solution, of course, is the one that uses the least number of steps to get to the goal.

There are two other aspects of search that should be mentioned here. In the scenarios just described, you began at your current state (no keys) and searched forward until you reached your goal (keys). This is called *forward search or forward chaining.* Suppose I showed you a diagram of your house with the location of your keys plainly marked, and you traced a path for yourself from the location of your keys (goal) back to where you are currently standing (current state). This is called *backward search or backward chaining.* People frequently use both of these search methods when trying to solve maze puzzles, either starting from "you are here" and tracing a path to the "goal box" or starting at the "goal box" and tracing backward to "you are here." You can do the same thing with lines of questioning or drawing inferences. Suppose you are a physician who is trying to diagnose a patient. You could ask the patient a number of questions such as "Do you feel faint?" or "When did the symptoms start?" and try to work your way forward to a diagnosis. Or you could start with a tentative diagnosis ("Hmm … there's a lot of flu going around; I wonder if that's the problem") and then work backward from the diagnosis, asking questions specific to that particular diagnosis, such as "Does your stomach hurt?" or "Do you have diarrhea?" Again, you could do either of these algorithmically (systematically investigate all symptoms at each stage of the diagnosis process) or heuristically (switching among symptoms classes and possible diagnoses depending on the patient's particular responses). Similarly, if you know your daughter has an interview and an exam scheduled for today, you start with that information to infer what mood she will be in when she comes home ("if she succeeds at both, she'll be happy"; "if she succeeds at only one, she'll be neutral"; "if she fails at both, we will have a total meltdown on our hands"), or you could take a look at her mood when she arrives home and infer the earlier outcomes ("she's happy so both must have gone well"; "she's neutral so one must have flopped"; "we've got a total meltdown, so both must have been flops").

Other heuristics offer varying degrees of success (such as deciding to get your girlfriend the same birthday present as last year since she liked that present so much last year) or estimating how many guys and girls were at a party you went to last week based on how many guys and girls you remember looking at. (You can figure out what's wrong with those.)

Finding the Way There

As Duncker pointed out, when we're unhappy with our current state and we have no immediate action at our disposal to reach our goal, we typically resort to *thinking*. The goal of that thinking process is figuring out which means (or actions) we can put into play to bring us closer to our goal. So the process of problem solving can be described as *difference reduction – employing actions that will reduce the difference between our current state and goal state*. This can be summarized in a heuristic called *means-ends analysis:*

Means-ends analysis: Heuristic problem-solving strategy

1. Analyze goal state
2. Analyze current state
3. Enumerate differences between the two states
4. Reduce the differences enumerated in No. 3 through
   - direct means
   - generating attainable subgoal states

Duncker brought the phenomenon of human problem solving into the laboratory for closer study. He did this by giving people a series of ill-defined problems to solve and then having them "think aloud" as they tried to solve them. An example of one of the problems he used is the now famous x-ray problem: A patient has a tumor in the middle of his abdomen. The tumor will kill the patient unless it is removed, but it is inoperable due to its proximity to vital organs. (Remember, this is 1945.) The tumor can be destroyed through the use of x-rays, but the dose needed also would destroy all of the fragile, healthy tissue surrounding the tumor, thereby killing the patient. How can the tumor be destroyed without damaging the healthy tissue or harming the patient? The participants' "think aloud" protocols showed systematic – and often ingenious – searches for means to remove or shrink the tumor while honoring the constraint of sparing the healthy tissue. A subset of these participants discovered a particularly brilliant solution called the convergence solution. (You can try to discover this solution before I discuss it later on.)

Means-ends analysis is not a problem-solving strategy that is restricted to humans. Instead, it is a natural strategy that can be observed in the behavior of other animals in natural habitats. The sophistication of the search process varies with the cognitive capacity of the organism studied. For example, orangutans are highly intelligent creatures who can often imitate what they see, particularly if the observed activity is to their liking. The orangutans of Borneo are notorious for pilfering canoes and cruising along streams in search of food, including the lunches stowed by nearby Camp Leakey personnel. One orangutan in particular, named Supinah, took a liking to washing clothes, an activity she observed among the camp staff along the river bank. But the staff were frightened of her and requested a guard to keep her away. This did not deter Supinah in the least. Here is a description of what ensued (reported by Byrne & Russon, 1998):

> Bypassing the guard meant detouring around him, which meant travelling through water because the end part of the dock where Supinah lurked stood knee-deep in water. Below this part of the dock was a dugout canoe. Supinah dealt with this situation … by freeing it and bailing it out … Riding the canoe required re-orienting it relative to the dock and raft, then propelling it alongside the dock towards the raft. Taking soap and laundry from the staff was then easy; Supinah merely hopped onto the raft, staff obligingly shrieked and jumped into the water, abandoning soap and laundry. Supinah immediately set to work washing the clothes …

Notice that Supinah not only searched for means to reduce the differences between her current state and her desired state (washing clothes), she generated a number of subgoals along the way, such as bailing out the canoe and re-orienting it in the direction she needed it to go.

Problems that require creating subgoals like this are typically difficult in large part because they tax short-term memory – you have to keep track of all of those subgoals, checking off the ones you've completed and maintaining those you haven't achieved yet. Which is why you can (e.g.) run an entire load of laundry sans detergent; the subgoal "add detergent" is lost from short-term memory along the way.

Artificial Intelligence: Machines That Think

Given Duncker's groundbreaking insights on problem solving, you might expect that his research spawned dozens of fruitful lines of psychological research. Not so. At the time Duncker was doing his research, psychology (particularly American psychology) was very much in the throes of behaviorism. The goal of behaviorism was discovering the laws by which behavior was shaped. Classical conditioning research showed us that the nervous system automatically learns complex contingencies between events so that they can be predicted (e.g., ringing bells means food is on the way). Instrumental conditioning research showed that behavior can be shaped by its consequences. But the most important thing was refraining from "speculating" about what was happening inside the "black box" (inside the head). The mind was off limits because internal states were unobservable – until the invention of computers, sophisticated symbol-manipulation machines that could solve problems and leave behind a printed trace of how they arrived at the solutions. Artificial intelligence, as it turned out, was chock full of internal states that could be readily interpreted as representations, search algorithms, and symbol manipulation. If machines could have internal states, why couldn't we?

Computer scientists Alan Newell and Herbert Simon in particular rediscovered Duncker's monograph on problem solving and wondered whether heuristics such as means-ends analysis could be automated. Could a computer be programmed to solve problems the way humans do – by seeking to reduce differences between start and goal states? The answer was an emphatic yes. In 1963, they published a paper describing their very fruitful development of artificial-intelligence programs that could prove theorems and solve problems. Their programs were called Logic Theorist and GPS (no, not "Global Positioning System" but "General Problem Solver"). Logic Theorist was developed specifically to do logical proofs, but the goal of GPS was far more ambitious – to solve a variety of problems using nothing more than means-ends analysis. Examples of the problems GPS could solve include first-order logic problems, cryptarithmetic problems, "Missionaries and Cannibals" problems, "Tower of Hanoi" problems, and mathematical integrals.

The groundbreaking programs led the way to the development of production systems, problem-solving programs that consist of the following components:

Condition-Actions Rules – Rules that specify which actions to take if a problem with a certain set of conditions is encountered
Long-Term Memory – Where condition-actions rules are stored
Working Memory – Where inputs are temporarily stored and matched to rules
Goal Stack – A "hidden" memory where subgoals are stored until they are completed
Conflict-Resolution Rules – Rules that resolve conflicts among the rules that are currently satisfied

For example, here are the rules for driving your car:

Rule 1: If engine is off, start engine.
Rule 2: If gas tank is low, fill tank with gasoline.
Rule 3: If raining, turn on windshield wipers.
Rule 4: If desired direction is forward, put gear into "drive" and accelerate slowly.
Rule 5: If desired direction is backward, put gear into "reverse" and accelerate slowly.
Rule 6: If desired direction is forward but forward path is blocked, put gear into "reverse" and accelerate slowly.

Suppose rule 4 and rule 6 both apply to the current situation. Both are demanding to be executed. Which rule should be applied? One conflict-resolution rule might be: *Choose the rule with greatest number of conditions.* So rule 6 would win because it has more conditions ("desired direction is forward" and "forward path is blocked").

Here is how a production system works:

1. Systematically compare inputs to the condition side of each rule.
2. When a match is found, execute the action stated in the rule.
3. If more than one rule matches, choose one or order the ones to be executed using conflict-resolution strategies.

This extremely simple but extremely powerful procedure will reduce the difference between current states and goal states, thereby solving problems the way humans do!

The ensuing decades saw the creation of hundreds of production systems. Some were used as research instruments to guide investigation into the limits and characteristics of human problem solving. Others were written for commercial application. These production systems were typically *expert system*s: Production-system programs that reproduce the performance of human experts.

In building such programs, programmers plumb the knowledge of experts in a particular domain, and then reproduce that knowledge in an automated production system. The most famous examples of medical expert systems are:

- MYCIN, an expert system for diagnosing and recommending treatment of bacterial infections of the blood, developed by Shortliffe and Buchanan at Stanford University
- Abdominal Pain System, an expert system for diagnosing acute abdominal pain, developed by deDombal at the University of Leeds
- Health Evaluation Through Logical Processing (HELP) System, a hospital-based system, developed at LDS Hospital in Salt Lake City that provides clinical, administrative, financial, and decision-support functions

Recent years have seen an enormous development in medical expert systems, including acute-care systems, decision-support systems, education systems, quality assurance, medical imaging, drug administration, and laboratory systems.

The advantages of these systems are many. They make expertise available to users who may not have access to experts (such as doctors in remote rural areas). They can hold and maintain significant levels of information. They provide consistent answers for repetitive tasks, and they work around the clock.

Expert systems, however, also have disadvantages. Most notably, they lack the common sense needed in some decision making. And they typically don't adapt to changing environments, unless the knowledge base is deliberately changed.

A painful example of this is the "flash crash" that occurred in the stock market on May 6, 2010. At 2:45 p.m. EST, the Dow Jones Industrial Average plunged 9% (about 900 points), and then recovered its losses miraculously within minutes. Trading had been turbulent that day because of concerns over a European debt crisis. According to an investigation by the U.S. Securities and Exchange Commission (2010), the financial firm Waddell and Reed then initiated a sell program that operated on an algorithm that had "no regard to price or time." This created a "hot potato" effect in which high-frequency traders started selling like mad, creating a "run" on the stock market. The situation was noticed and corrected within minutes. But it left behind a distrust of relying too implicitly on automated systems for high-level decision making.

How Experts Solve Problems

Who corrects the automated problem solvers? The same people whose knowledge and skills were plumbed in order to create them: experts. So what makes someone an expert problem solver? The short answer is this: Becoming an expert means increasing your knowledge in a particular domain (expertise is domain-specific) and organizing your knowledge efficiently so that solutions can be retrieved or constructed quickly. The importance of the latter cannot be overestimated. The larger the knowledge base, the more information there is to search through, and the longer the search should take. But expert knowledge bases are typically organized around problem-relevant (or solution-relevant) features. Hence, solution retrieval or construction is blindingly fast!

Consider, for example, chess grandmasters. Chess is a complex game that requires strategy and thinking ahead. For each possible configuration of chess pieces at each stage of the game, there are numerous possible moves. Some of these are more strategic than others; they allow greater control of the board and greater likelihood of achieving checkmate. When considering each move, you must also consider the moves available to your opponent in response, which moves are then available to you in response to each of your opponent's possible moves, and so on. The number of legal moves for six-ply-deep "look ahead" is approximately 1.8 billion; yet grandmasters rarely look farther ahead than 100 branches in 15 minutes (Gobet & Simon, 1996). And they still regularly beat automated chess programs that engage in large-scale algorithmic searches. In fact, a chess program beat a grandmaster only once – and the outcome of that tournament is still disputed. On May 11, 1997, IBM's chess program *Deep Blue* beat reigning world champion Garry Kasparov in a six-game tournament (two wins for Deep Blue, one for Kasparov, and three draws) (Newborn, 1997). After the loss, Kasparov essentially accused IBM of cheating. He said that human assistance must have underlie the deep intelligence and creativity in the machine's moves, and this kind of intervention was forbidden by tournament rules. IBM denied the charges, saying the only human intervention occurred between games. Kasparov challenged *Deep Blue* to another tournament on national TV, but IBM refused and instead chose to retire *Deep Blue.* So what might this "human intervention" have been like?

William Chase and Herbert Simon (1973) studied mental characteristics of chess players at varying levels of expertise, novice, intermediate, grandmaster (Simon himself was a grandmaster). They found no differences in knowledge of chess rules, overall intelligence (IQ), memory span (number of items that can be held in short-term memory, or number of moves in look ahead). They did, however, find sizeable differences in terms of how these players "parsed" chessboard positions – that is, how well they perceived and remembered strategic chess configurations. They found this out by showing diagrams of boards of mid-game configurations or random configurations for five seconds and then asking the players to reproduce what they saw using a real game board and real chess pieces.

When the configurations were *random* (could not actually occur in a real game), they found no difference between experts and novices. But when the configurations were real ones that could actually occur in actual games, experts were able to replace more pieces in correct configurations, required less time to reconfigure the boards from memory, and reproduced the configurations strategically (e.g., reproducing a castled king configuration, followed by a fianchettoed bishop configuration, and so on). Chase and Simon interpreted these results to mean that the chess experts' knowledge (memory) bases were organized around strategic chess patterns, along with associated optimal moves. Because their memories were so strategically organized, experts were quicker to visually "parse" a game board into patterns of strategic attacks (e.g., queen's bishop attacking rook, which is guarded by queen's pawn), and quicker to retrieve optimal responses to these patterns.

More recently, Reingold and colleagues (2001) posed a more challenging task to chess novice, intermediate, and master chess players, and found results that corresponded with Chase and Simon's. Players were shown diagrams of chessboards for one second, and then the position of one piece was changed. The original and the modified diagram were displayed alternately for 1 second with a 100-millisecond blank between the displays. This was continued until the player could identify the change. Once again, there was a difference among the groups on the random-configuration displays. But experts were on average about one second faster to detect the change than novices or intermediate players. Moreover, when the configurations were random, all players were found to be able to scan about 10 squares during a given exposure. But when the configurations were real, experts were found to scan an area of about 25 squares – more than twice the size!

The pieces scanned also differed among the groups. Experts were found to visually scan strategic pieces more frequently than non-strategic and to scan empty squares between strategic pieces (i.e., places where pieces can move to attack or defend other pieces). To an expert, empty spaces weren't irrelevant; instead, they were lines of possible strategic attack. It turns out that expert soccer players do the same thing: They watch the movements of players who do *not* have the ball but could receive it via a pass. Novices, on the other hand, tend to look only at the ball and the player who is controlling it (Williams and colleagues, 1992 and 1994).

Similarly, novice and expert physicists respond quite differently to physics-based problems. In a classic series of studies, Chi, Feltovich, and Glaser (1981) asked physics graduate students and working physicists to sort and solve a number of physics problems taken from standard textbooks. The graduate students had sufficient knowledge of physics to solve these problems. But they approached them very differently than did the working physicists. When asked to sort the problems, the graduate students sorted them according to similarities in surface features. For example, they classified together problems that referred to inclined planes, putting those that referred to springs in a different category. The working physicists, on the other hand, sorted the problems according to the underlying physics principles needed to solve the problems. Not surprisingly, the physicists were more likely than the students to find a solution to the problems and more quickly.

These results indicate that experts really do "see" problems differently from novices, in large part because their efficiently organized memories direct their attention to solution-relevant features of problem situations. Because they're looking in the right place at the right things, constructing a solution is immensely easier. And that is how a grandmaster can play dozens of novices simultaneously (and win), Beckham can be exactly where he needs to be to intercept a soccer ball, and Einstein or Feynman can find solutions to physics problems that appear intractable to others.

Insight and Genius

So far, I've shown how problem solving – even expert problem solving – can be analyzed as a search process. But what about insight, that seemingly magical type of problem solving that seems to appear out of nowhere with a sudden gasp of "eureka?" Surely, this must be a different kind of animal, right? The answer is yes – and no. There are certain characteristics of insight that do not apply to "normal" problem solving – particularly the "aha!" recognition of having reached a sudden solution. But a closer look by neuroscientists into this phenomenon suggests that it really may reduce to search processes after all.

While Gestalt psychologist Karl Duncker gave us a precise way to describe "regular" problem solving, another Gestalt psychologist, Wolfgang Köhler (1887–1967), transformed the way we think about insight. Köhler, a psychologist trained at the University of Berlin, was working at a primate-research facility maintained by the Prussian Academy of Sciences in the Canary Islands when the First World War broke out.

Marooned there, he had at his disposal a large outdoor pen and nine chimpanzees of various ages. Behaviorism, which was the dominant theoretical framework at the time, described problem solving as an incremental process involving trial and error during which behaviors that led to satisfying outcomes increased in frequency whereas those that led to unsatisfying outcomes reduced in frequency. But Köhler's chimpanzees seemed to be displaying a very different kind of problem solving; solutions seemed to occur suddenly and all at once, not through incremental trial and error. When he hung bananas from a wire that was too high for his chimps to reach, they jumped and meandered about trying unsuccessfully to reach it. Then, suddenly, they stacked nearby boxes and climbed on them to reach the banana. Sometimes they even cooperated in building the makeshift platform. His most intelligent chimp, named Sultan, solved a number of problems this way. The most notable was fitting sticks together to reach a banana that was placed outside his cage and out of reach. Sultan had tried to reach the banana with each stick, but they were all too short to reach it. While fiddling with them, he happened to fit one inside the other, thereby lengthening its reach. He then immediately set to work retrieving the banana, indicating that he *immediately* realized the implication of his accidental discovery.

Köhler described the kind of problem solving he observed as "insight." More precisely, he defined insight as an *sudden understanding of the relevant relations among elements of a problem*. This sudden flash of understanding frequently yields a restructuring of the problem. And this is the key: Most creative solutions typically involve spontaneous restructuring of problems features in ways that are atypical for that type of problem.

Why and when do these insights occur? Two explanations have been offered to answer this question. The first is referred to as the *special-process explanation: Insightful solutions happen instantaneously.* The problem solver changes suddenly from a state of complete ignorance to a state of complete knowing. The second explanation is referred to as *the business-as-usual explanation*: *All problem solving is an incremental process in which one gradually gets closer and closer to a problem solution.*

In the 1980s, Janet Metcalf and colleagues (1986 and 1987) systematically investigated "normal" and "insight" problem solving in rather ingenious ways. They used three different types of problems, and all participants had to solve problems of each type: Incremental problems that could be solved with college-level familiar algorithms (such as solving 3 × 2 + 2x + 10), incremental problems that could be solved using simple heuristic search (such as, given containers of 163, 14, 25, and 11 ounces, and a source of unlimited water, obtain exactly 77 ounces of water), and insight problems, such as this: *Water lilies double in area every 24 hours. At the beginning of summer there is one water lily on the lake. It takes 60 days for the lake to become completely covered with water lilies. On which day is the lake half covered?*

Every 15 seconds, the participants were required to give "warmth ratings" of how close they felt they were to a solution using a seven-point scale where 1 corresponded to cold ("I'm clueless") and 7 corresponded to hot ("I'm sure I've got the answer").

The results seemed to support the "special-process" explanation: Warmth ratings for the non-insight problems showed a slow and steady increase from 1 to 7 as participants worked on their solutions. For insight problems, however, participants consistently gave ratings of 1 or 2 up to the point when the solution occurred to them, at which point, they gave ratings of 7. In other words, they immediately switched from a state of complete cluelessness to one of complete understanding.

In 1990, Bowers and colleagues took a slightly different tack using Compound Remote Associates (CRA) problems. These problems consist of trios of words. For each trio, the solver's task is to find a fourth word that can be joined to the beginning or end of all three words in the set to form a real word. Here are some examples:

a. *french, car, shoe*
b. *boot, summer, ground*
c. *table, wall, dog*

The first two trios actually have solutions (*horn* and *camp*, respectively). The third trio is unsolvable; the words have nothing in common. Participants were shown pairs of CRA trios, where one was solvable and one was not (e.g., *a* above paired with *c*). Each pair was presented side by side for a few seconds during which participants had to give (a) the solution to the coherent triad (when they could), (b) make a forced-choice decision regarding which of the two triads was solvable (even if they hadn't solved either of them), and (c) a confidence rating from 0 (no confidence) to 2 (high confidence) that their forced-choice decision was correct.

Interestingly, even when participants could not actually find the solution to trios, they could reliably guess which trios actually had solutions and which did not. This is problematic for the "special-process" explanation of insight: If insightful solutions occur in a sudden, "all-or-none" manner, then participants should have performed no better than dictated by chance when deciding which trios had solutions and which did not. After all, when they hadn't solved either of the trios, they should have been in a "clueless" state for both. So when asked which was solvable, they should have just guessed and been right about 50% of the time. But they didn't. On average, they selected the solvable CRAs over the unsolvable ones about 60% of the time. This is statistically higher than chance. So perhaps there was something incremental going on after all, just outside of awareness.

Here is where neuroscientific techniques developed in the last two decades come to the rescue. Bowden and colleagues (2005) had people solve CRA problems and non-insight problems while having their neural activity recorded via functional magnetic resonance imaging (fMRI) or event-related potentials (ERP). These techniques allow the researcher to see which areas of the brain are active during problem solving. What they discovered changed the way we think about insight.

They found more right-brain activity for insight problems. But more interestingly, they found an abrupt switch from left-hemisphere activity to right-hemisphere activity just prior to solutions on insight problems. This abrupt change was not observed with non-insight problems. So far, so good for the special-process explanation. But now here are the results that show insight has a vital "business-as-usual" component as well.

The whole problem-solving process for insight problems looked like this. First, there was strong activation in the dominant (left)

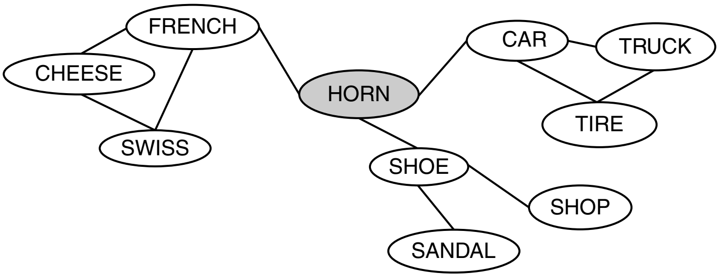


FIGURE 8.5. How memory is automatically searched, sometimes without awareness.

hemisphere and weak activation in the non-dominant (right) hemisphere. This means that the dominant meanings of the word trios were strongly activated. The problem is that these dominant meanings aren't particularly helpful. The dominant meaning for "French" (i.e., a language spoken in France) has very little to do with the dominant meaning for "car" (i.e., transport vehicle for the average person) or "shoe" (i.e., protective covering for your foot). But there are non-dominant meanings for these words as well, and they do have something in common (i.e., "French horn," "car horn," and "shoe horn"). These non-dominant meanings are activated as well, but much less strongly. Over the course of a few seconds, this weaker activation continues to spread and grows in strength, particularly if the dominant meanings are allowed to diminish in strength (i.e., you stop thinking about them incessantly). Figure 8.5 shows how it might look.

When the non-dominant meanings are sufficiently activated, they "spring" into consciousness, and the solver suddenly recognizes the solution. This is seen as an increase in alpha-band activity elsewhere in the brain about 1.5 seconds prior to a gamma burst that accompanied the solution to the problem. So these solvers were actually working on the problem outside of conscious awareness. They only became aware of this when the solution suddenly sprang into consciousness. But, at bottom, they were searching memory for a solution all the while. So insight seems to be a particular blend of both "business-as-usual" search and "special-process" sudden discovery.

If insight is a simple incremental process at bottom, why does it happen so rarely? On average, people produce insightful solutions only about 25% of the time. The answer seems to be that insight is rare because insight (like other aspects of cognition) suffers from framing effects: *The way a problem is described influences how we think about it.* This limits the types of solutions that occur to us if the solution requires seeing the problem in a different light.

The Einstellung effect is a perfect example of this – and it explains why a novice can sometimes find a solution to a problem that experts have failed to find. *Einstellung refers to the negative effect of previous experience when solving new problems.* It has been demonstrated in a number of different contexts, but the most famous examples come from the work of yet another Gestalt psychologist: Abraham Luchins (1942, 1959). Luchins asked people to solve "water jug problems," which require the solver to figure out the most efficient way to measure out a precise amount of water using (usually) three jugs of different capacities. Some examples are given in Box 8.2. Try working through the problems, beginning with the first and continuing until you solve the last one.

Box 8.2. Luchins Water Jug Problems

Problem	Capacity of Jug A	Capacity of Jug B	Capacity of Jug C	Desired Quantity
1	21	127	3	100
2	14	163	25	99
3	18	43	10	3
4	9	42	6	21
5	20	59	4	31
6	23	49	3	20
7	15	39	3	18
8	**28**	**76**	**3**	**25**
9	18	48	4	22
10	18	48	4	22

Solutions:

All problems except 8 can be solved by B – 2C – A. For problems 1 through 5 this solution is simplest, but notice that problems 7 and 9 can

(*continued*)

be solved using a simpler formula (A + C), and problems 6 and 10 can also be solved using a simpler formula (A – C). Problem 8 cannot be solved by B – 2C – A, but can be solved by A – C.

Did people notice these simpler formulas? Nope. Of Luchins' participants, 83% used B – 2C – A on problems 6 and 7. They had gotten into a rut and just didn't see the simpler solution. The failure rate for problem 8 was 64% because the participants tried to use the formula they had become used to applying, and when it didn't work, they couldn't find the simpler one. On problems 9 and 10, 79% used B – 2C – A, even though a simpler formula could have been used.

When Luchins gave people only the last five problems, less than 1% used B – 2C – A, and only 5% failed to solve problem 8.

When he warned them "Don't be blind" after problem 5, more than 50% were able to find the simpler solution on the remaining problems.

Now, here is the interesting thing: The first five problems can all be solved using the same (somewhat complicated) formula. The remaining problems can be solved using much simpler formulas. But by the time people got to problem 6, they were "stuck in a groove"; they continued to try to solve the remaining problems with the same formula. This meant that problem 8 couldn't be solved at all, and the others could be solved in a single step. Their experience solving the first five problems interfered with their ability to solve the remaining problems efficiently – or in the case of problem 8, finding a solution at all.

One way to think about this is to take another look at the search illustration in Figure 8.3. Suppose you did a depth-first search, and you're down there in the basement far away from the goal with no way to get there. The best thing to do is to go back up to the top and start again, give yourself a fresh start. Even better, give yourself a fresh start after a good rest.

This is sometimes referred to as the incubation effect, when interrupting a task actually improves success rate. Consider this problem used by Silveira (1971):

> *You are given four separate pieces of chain that are each three links in length. It costs 2 cents to open a link and 3 cents to close a link. All links are closed at*

> *the beginning of the problem. Your goal is to join all 12 links of chain into a single circle at a cost of no more than 15 cents.*

One group worked on the problem for half an hour, and only 55% of the people solved it. Another group worked for half an hour and then took a half-hour break in which other activities were performed. The success rate among that group was 64%. A third group also took a break, but their break lasted four hours. Their success rate was 85%!

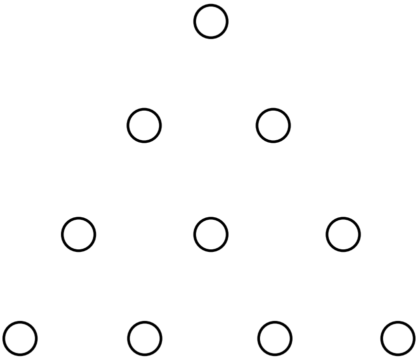
So let's go back to Duncker's x-ray problem. Have you found a solution? If not, think about the following story: A physicist is conducting an experiment using a specially constructed, very expensive, and very fragile light bulb. The filament in the light bulb breaks but can be repaired with heat delivered by a laser. The laser can deliver light at different wavelengths that yield different temperatures. The problem is that the temperature required to fuse the filament would melt the glass surrounding it. The physicist thinks about this for a while and then retrieves several lasers from adjoining labs. She positions the lasers so that they surround the bulb, and when they are all turned on, the light beams converge on the filament at the same time. She sets all of the lasers to low-heat intensity and turns them on. When the beams converge, their temperatures sum and quickly reach the high temperature needed to fuse the bulb. Now do you know how to save Duncker's patient? If not, continue onto the next chapter, where I'll discuss problem solving through analogy.

And if you want to try your hand at more insight problems, try your hand at the ones in Box 8.3 (from Metcalfe & Wiebe, 1987).

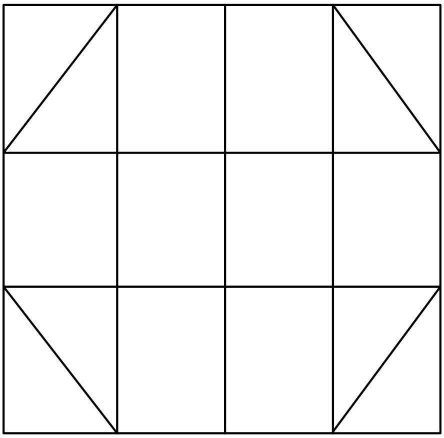
Box 8.3. Try Your Hand at These Insight Problems

1. Describe how to cut a hole that is big enough for you to put your head through in a 3 × 5 inch card.
2. The triangle shown below points to the top of the page. Show how you can move three circles to get the triangle to point to the bottom of the page.

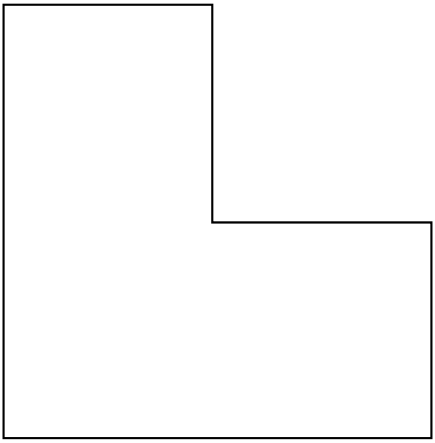
(*continued*)



3. A man bought a horse for $60 and sold it for $70. Then he bought it back for $80 and sold it for $90. How much did he make in the horse-trading business?
4. A woman has four pieces of chain. Each piece is made up of three links. She wants to join the pieces into a single closed ring of chain. To open a link costs 2 cents and to close a link costs 3 cents. She has only 15 cents. How does she do it?
5. A landscape gardener is given instructions to plant four special trees so that each one is exactly the same distance from each of the others. How would you arrange the trees?
6. A small bowl of oil and a small bowl of vinegar are placed side by side. You take a spoonful of the oil and stir it casually into the vinegar. You then take a spoonful of this mixture and put it back in the bowl of oil. Which of the two bowls is more contaminated?



7. How can you draw the figure above without raising your pencil from the paper, without folding the paper, and without retracing any lines?
8. Describe how to put 27 animals in four pens in such a way that there is an odd number of animals in each pen.
9. Show how you can divide the figure below into four equal parts that are the same size and shape.



10. Show how you can arrange ten pennies so that you have five rows (lines) of four pennies in each row.

NINE

Analogical Reasoning

THIS IS LIKE THAT

In the annals of financial history, the year 2008 stands out like a tarantula on white bread. That was the year the banking industry faced its worst crisis since the Great Depression. Unprecedented rises in real-estate prices during the previous decade seduced bankers into making riskier and riskier mortgage loans. When the housing bubble burst, so did their mortgage portfolios. The banking behemoth Lehman Brothers went bankrupt. Others, such as Merrill Lynch and AIG, came within a hair's breadth of failing as well until the federal government stepped in to rescue banks deemed "too big to fail."

This was an enormous and potentially unpopular undertaking, and so Federal Reserve Chief Ben Bernanke appeared on the TV news show *60 Minutes* to explain to the county why we needed to bail out the banking system. He did so by way of analogy. Imagine, he explained, that you have an irresponsible neighbor who smokes in bed and sets fire to his house. Should you call the fire department, or should you simply walk away and let him face the consequences of his actions? What if your house – indeed the houses in the entire neighborhood – are also made of wood? We all agree, he argued, that under those circumstances, we should focus on putting out the fire first. Then we can turn to the issues of assigning blame or punishment, rewriting the fire code, and putting fail-safes in place.

This is not a chapter about the merits or failings of the banking-industry bailout. It is about the power of analogy to explain and persuade. A powerful analogy can deliver more "bang for the buck" than hours of lecture, dozens of charts, or lengthy documentaries. Bernanke's "fire in the neighborhood" analogy was ingenious because it was easy for the average viewer to apply what they understood about fires (they can spread and destroy all of our houses) to the financial crisis (if these mega-banks fail, our money and credit-lending opportunities will disappear as well). But even strong analogies can backfire. Why? Because they usually carry excess baggage that can just as readily support other interpretations.

As it turned out, numerous financial and political analysts didn't buy Bernanke's analogy. In the following weeks, so many people skewered it that it ended up looking like Swiss cheese in a New York deli. Google "Bernanke burning analogy," and you will get 181,000 hits. Here is a particularly lucid analysis of the flaws in this analogy, written by Professor Michael Hudson and posted on the Centre for Research on Globalization Web site. (http://informationclearinghouse.info/article22229.htm)

> What's false about this analogy? For starters, banking houses are not in the same neighborhood where most people live. They're the castle on the hill, lording it over the town below. They can burn down and leave the hilltop to revert "back to nature" rather than having the whole town gaze up at a temple of money that keeps them in debt. More to the point is the false analogy with U.S. policy. In effect, the Treasury and Fed are not "putting out a fire." They're taking over houses that have not burned down, throwing out their homeowners and occupants, and turning the property over to the culprits who "burned down their own house." The government is not playing the role of fireman. "Putting out the fire" would be writing off the debts of the economy – the debts that are "burning it down."

Why was it so easy for viewers to understand the "burning house" analogy and so easy for critics to knock it down with other, just-as-easy-to-understand analogies? Cognitive scientists Keith Holyoak and Paul Thagard (1989) claim that the "analogical mind is simply the mind of a normal human being." Fellow cognitive scientist and Pulitzer Prize winner Douglas Hofstadter (2009) describes analogy as "the core of cognition." Developmental psychologists have frequently noted that young children develop this capacity on their own without any direct or specialized instruction from parents or teachers. It just seems to be the way the mind works. This chapter explains why.

Analogy as It Should Be Done

An analogy is a relational similarity. Analogies are objects or events that are isomorphic; they have structures in common. In other words, two objects or events are not considered analogs because they refer to the same things but because the relationships among those things are the same. A burning house has nothing physically in common with a failing bank. But at a deeper level of analysis, they describe similar relations: A burning house is in danger of physical collapse. A failing bank is in danger of financial collapse. More importantly, both require intervention if the collapse is to be prevented.

We could say, then, that the *surface features* of the situations may differ, but the *relationships among the features* are the same. The *target* of analogical reasoning is the thing we are trying to understand (e.g., the financial bailout). The *base* is the object or event to which the target is compared (e.g., burning house). *When we engage in analogical reasoning, we map the relational features from the target onto the base so that they correspond to one another.* The result is that we understand the target in the same way that we understand the base. Using Bernanke's analogy, "house" corresponds to "bank," "burning" corresponds to "assets losing value," and "neighborhood" corresponds to "your bank and loan accounts." The solution to the burning-house scenario is pouring water onto the fire. So, by *analogical transfer*, the solution to the failing-bank scenario is pouring money into the banking industry. The water stops the fire, and the bailout money stops the asset accounts from losing value.

As this example shows, the power of analogical reasoning is that it allows information to be transferred from the base to the target, enhancing our understanding and, frequently, suggesting a possible solution to the target problem. If we were to strip away the surface features of these analogous situations, we would end up with an abstract schema for solving these kinds of problems that might go something like this: Whenever a vital resource is in danger of being depleted, pour something onto it that will stop the depletion. Notice that this schema makes no mention of banks, houses, fires, assets losing value, water, or money. But it applies to both scenarios, and more importantly, any scenario that shares those relational structures.

The problem with relying on analogies to understand and solve problems is that analogies are a little like bulls running free in a china shop. Without proper constraints, they can cause a lot of damage, leading to a bigger mess than you started with. As we saw in the previous chapter, thinking frequently falls prey to Einstellung – previous experience can hinder our ability to see a new, easier, or any solution to a current problem. Why? Because if the similarity between current and past problem triggers reasoning by analogy, you will automatically try to use same-solution strategy – even if it doesn't work or makes you work harder than you really need to work. Unfortunately, decades of research have uncovered no hard and fast rules (algorithms) for constraining analogical inference. What we have discovered instead are heuristics and principles that frequently prove useful.

Consider a simple example that will take you right back to high school science. Your teacher probably introduced the Rutherford model of the atom by saying this: *An atom is like the solar system*. The target (the thing she was trying to get you to understand) was the atom. The base (the thing she assumed you already knew) was the solar system. She may have continued with this: *The nucleus of the atom is like the sun, and the electrons of the atom are like the planets.* This statement invites you to map the base feature "sun" onto the target feature "nucleus" and the base feature "planets" onto the target feature "electrons." Finally, she stated: *Electrons orbit the nucleus the way planets orbit the sun*. So far so good. Now what inferences could we draw by analogy? How about this? *The sun is more massive than a planet, so the nucleus is more massive than an electron*. OK, that works. How about this? *The sun is hotter than a planet, so the nucleus is hotter than an electron*. No, that one is right out. But why? What principle (or set of principles) allows the first two inferences but excludes the third?

In 1983, Dedre Gentner proposed a Structure Mapping Theory of analogical reasoning the core components of which consist of two constraining principles. The first was the *relation-mapping principle: Prefer analogs with similar relations over those with similar attributes.* ("Attributes" is another term for "surface feature.") The second was the *systematicity principle: Prefer analogs that allow coherent systems of relations to be mapped rather than individual relations.* This means we should prefer predicates that link relation terms over predicates that simply refer to individual attributes. Let's see how these work.

First, let's represent the information we've got so far in terms of relation terms (predicates) and attributes, like this:

Mass-of (sun) Mass-of (nucleus)
Mass-of (planet) Mass-of (electron)
Orbits (planet, sun) Orbits (electron, nucleus)
Greater-than [Mass-of (sun), Mass-of (planet)]
Is-hot (sun)

Using the *relational principle*, we see that this is a pretty good analogy: The attributes don't match ("sun" and "planet" are on one side, "nucleus" and "electron" are on the other), but all of the relational terms do match ("mass-of" and "orbits"). Now let's use the *systematicity principle* to draw inferences. This principle says we should prefer to transfer higher-order relations. "Is hot" is an individual relation; it refers only to a single attribute and doesn't link other relations together. So it's a bad choice to transfer from base to target. But "greater-than" links two relations that appear in both target and base. So we should feel free to transfer that relationship. That makes "greater-than [mass-of (nucleus), mass-of (electron)]" a justifiable inference.

Gentner's Structure Mapping Theory was extended by Keith Holyoak and Paul Thagard (1997) in a *multi-constraint theory* that measures the coherence of an analogy on three dimensions: structural consistency, semantic similarity, and purpose. Essentially, this theory favors analogies that have structural consistency along the lines of Gentner's systematicity principle (structures of related higher-order relations are preferred), but it also measures analogies based on similarity in meaning of the relations and attributes referred to in the analogs, along with the reasoner's purpose for engaging the analogy process. These multiple constraints help ensure that analogies will be chosen and mapped in realistically meaningful ways that facilitate reaching a particular goal.

Analogy: How It Is Actually Done

So far, we've learned what researchers have discovered about optimal analogies. But is this the way people actually use them to solve problems, make decisions, or persuade others?

Let's look first at lawyers in the courtroom. Lawyers must persuade juries and judges by making cogent arguments based on sound legal reasoning. It may surprise you to learn that analogy plays a vital role in this process. In law, analogies are frequently used to draw parallels between an undecided case and cases that have already been decided. Attorneys and judges will apply the decision from a previous case to one currently under consideration if they believe there exist sufficient similarities between the two. If the cases match up perfectly, the argument is referred to as "on point" (or, in common parlance, a slam dunk). Since lawyers and judges are specifically trained in this kind of cogent reasoning, they surely constitute expert analogical reasoners. The level of performance they achieve is surely the best we can expect.

A good example of this is *Adams v. New Jersey Steamboat Company,* 151 New York, 163 (1896). Adams was a steamboat passenger on an overnight trip from New York to Albany. While he slept, his money was stolen, despite his having locked his door and his windows. The case turned on this crucial point: Is a steamboat more like a "floating hotel" or a "railroad on water?" If it is more like a hotel, then the defendant was liable as an insurer because this is the rule for innkeepers. If it is more like a "railroad on water," then the steamboat company was *not* liable because railroads are *not* insurers of their passengers' property. They are liable only if negligence on their part can be demonstrated.

The court decided that a steamboat was more like a "floating hotel" than a "railroad on water": Hotels have rooms to accommodate guests, and steamboats have rooms to accommodate passengers. Hotel guests and steamboat passengers pay for the use of rooms for the same reasons – so they have a private place to sleep. As a result, both are vulnerable to the same risk of fraud and plunder, which creates a "special relationship" between hotel and guest on the one hand and steamboat and passenger on the other. This "special relationship" carries with it responsibility for the guest or passenger on the part of the hotel or steamboat. As far as the court was concerned, the analogy between the two was so strong that "The two relations, if not identical, bear such a close analogy to each other that the same rule of responsibility should apply." So the court applied the rule for hotels to steamboats.

Moreover they also considered and rejected the counter-analogy that a steamboat was like a "railroad on water" due to a crucial difference: The steamboat operator takes entire charge of the traveler by assigning a private room for exclusive use. Railroads don't do this. They provide limited sleeping accommodations to a few passengers in a car with open berths.

As this example shows, the court's decision depended in large part on which analog was judged to best match the undecided case. As is also clear, the match was considered on both surface and structural levels. Precedents involving hotels and railroads were brought to because they overlap significantly in surface features. But these surface features were not the deciding factor. But the case was decided on the basis of structural similarity, not surface similarity. Assigning a private room for exclusive use implies assumption of greater liability on the part of the company than the assigning of a berth in an open car where the opportunity for theft is less preventable. Perhaps the larger point is the lawyer who makes the strongest analogy wins the case.

The strategy behind analogical reasoning is perhaps best summarized this way: When you are not sure how to solve a problem, find a similar problem that you know how to solve and apply that solution. More precisely, we can say that analogical reasoning involves:

1. Searching memory for a similar problem that you know how to solve.
2. Mapping the structural correspondences between new and old problems.
3. Applying the solution to new problem.

Which is the most difficult step? Believe it or not, it is step 1. Oh, people do indeed automatically retrieve previous problems or cases based on similarity. But the cases retrieved tend to be similar in terms of surface features, which aren't particularly useful for deriving solutions. In fact, they can mislead you into applying completely inappropriate solutions to the current situation.

In the previous chapter, you read about the Duncker x-ray problem: A patient has an inoperable tumor in his abdomen that can be destroyed with radiation. But the high dose required to destroy the tumor would also destroy the healthy tissue surrounding the tumor, thereby killing the patient. You probably found this problem difficult to solve. But then you read about a physicist who needed to use a laser to fix the filament of a light bulb, but the laser intensity needed would shatter the fragile surrounding glass bulb. So she surrounded the light bulb with lasers and set each on a low intensity such that they converged on the filament at the same time, summing in intensity and fusing the filament. You probably noticed the similarity between these two problems and applied the convergence solution to the x-ray problem: Surround the patient with x-ray machines that each deliver a low dose of radiation such that they converge on the tumor at the same time.

What if I had given you the following story as a hint rather than the light-bulb story? Do you think you would have noticed the similarities between it and the x-ray story?

> *A fortress surrounded by a moat is connected to land by numerous narrow bridges. An attacking army successfully captures the fortress by sending only a few soldiers across each bridge, converging upon it simultaneously.*

The answer is probably not. Only 10% of people solve the x-ray problem using the convergence solution when the problem is presented alone. When they are shown the fortress story first, the success rate improves, but not by much; only 30% of people successfully see the similarity between the problems and apply the convergence solution (Gick & Holyoak, 1980). But when they are shown the light-bulb story first, the results are much more dramatic; 70% of people spontaneously see the similarity between the x-ray and light-bulb problem and successfully apply the convergence solution (Holyoak & Koh, 1987).

Why the difference? Because memory retrieval is strongly driven by surface-feature similarity, especially among novices. Despite the fact that the x-ray, light-bulb, and fortress problems all have identical problem structures, the similarity between their surface features differs greatly. People find x-rays to be more similar to lasers than to armies. They also find fragile tissue to be more similar to fragile glass than to narrow bridges. When reading about x-rays and fragile tissue, people are more likely to be reminded of a previous story about lasers and fragile glass than of a previous story about armies and bridges.

Ross (1984) demonstrated this quite dramatically in a set of studies on the use of examples in learning. People were given three pairs of mathematical formulas to learn, such as permutations and combinations. These formulas were taught by giving them examples. The examples consisted of word problems describing different topics, such as pizza delivery, a drunk finding his keys, and so on. Then the learners were given a test containing word problems with the same topics and formulas but paired differently than during the learning phase. For example, suppose the instructive example for permutation was a story about a drunk finding his keys, and the instructive example for combination was a guy delivering pizzas. At test time, these were switched so that the story about a drunk was actually a combination problem and the test story about pizza delivery was really a permutation problem. Participants were instructed to "think aloud" as they solved the problems so that memory retrievals could be tracked. The question was, which aspect of the problem would cue memory? When trying to solve the new pizza-delivery story, would they say "Oh, another pizza problem" or "Oh, another permutation problem?" Which aspect of the problems would cue memory retrieval?

The answer was definitive: More than 70% of "remindings" were story-content driven. Not surprisingly, if subjects said, "Oh, another pizza problem" instead of "Oh, another permutation problem," they would get the wrong answer because they would try to solve the new pizza problem using the combination formula that was associated with pizzas during instruction.

Why do novices retrieve problem analogs based on surface-feature similarities? Because the success of your search for a problem analog will only be as good as the quality of the database (knowledge base) you're searching. When you're learning a new domain, you don't really know what is important and what isn't. Surface features get stored right along with the important stuff and are oftentimes more salient (more vivid or understandable) than the important content you are supposed to learn. This means that becoming an expert in a field depends in large part on restructuring your knowledge base to abstract away from irrelevant surface details in favor of problem structure (relational) information.

You would think that the best way to improve your chances of retrieving a good analogy would depend on experience – solving a lot of problems in the domain of interest. That turns out to be only part of the story. The crucial aspect is not simply solving the problems but *comparing* them.

Gick and Holyoak (1980) were able to increase solution success on the Duncker x-ray problem by having people read more than one analogue (e.g., the light-bulb and the fortress story) and specifically describe how they were similar. Under these conditions, success rate soared to higher than 75%. Kurtz & Loewenstein (2007) systematically studied the impact of solving or comparing multiple problems prior to or at the same time as solving the x-ray problem and found the same thing: People were far more likely to see the structural similarity between the x-ray problem and other problems – and to solve the x-ray problem using the convergence solution – if they were required to explicitly compare problems.

Cummins (1992) demonstrated the power of comparing relational structures on solution success even more directly. In this series of studies, people were required to read algebra word problems, to sort them into solution-relevant categories, and to solve them. Some people just read the stories, sorted, and solved them. Another group had their attention drawn to important mathematical relations in the problems when they were required to verify the information (e.g., *it took 10 hours to make the trip – yes or no?*). A third group was required to compare the same mathematical relations across problems (e.g., *10 hours is to completing the trip as 3 hours is to filling the vat – yes or no?*) When asked to sort the problems into solution-relevant categories, those who simply read the problems or verified information sorted them according to similarities in surface features. But those who were required to compare relational information across problems sorted them according to mathematical structure. The latter group also was more successful in finding solutions to the problems.

The results of this work make it clear that comparing cases allows people to abstract away from irrelevant details and remember solution-relevant relational information, whether that means mathematical structure or legal structure. If memory is organized this way, the likelihood of retrieving a useful analog from memory when trying to solve a new problem is much greater.

Why Analogy Is the Core of Cognition

Psychological researchers who study cognition are like precocious, inquisitive kids who take a watch apart to see how it works. They carefully unpack each component, testing it alone and in combination with other components to see how it works and what it does. In the end, they understand what role each of these components plays in creating a whole object that keeps time. The next step is to gather up all of the components spread across the floor and put them back together to see if they understand how to build a functioning watch. If they haven't got it right, they end up with a piece of abstract art rather than a functional object that tells time.

Since cognition is information processing, building a cognitive system is more like building a digital watch than a gears-and-hands watch. You need to have the right kind of solid-state circuitry and the right kind of programming to make it work. Early models of analogical reasoning were production systems consisting of rules and separate registers for long-term memory, working memory, and subgoal structures. The components of analogy were coded as rules. The rules were systematically compared to inputs, conflicts among matching rules were resolved, and the steps of analogical reasoning were churned out. One of the first of these was the Structure Mapping Engine, developed by Falkenhainer, Forbus, and Gentner (1989). This production system codified as rules the principles described in Gentner's Structure Mapping Model. Subsequent systems included Ferguson's

MAGI (1994), Kuehne's SEQL (2000), and Burstein's CARL (1986). Each of these met with varying degrees of success in reproducing true analogical inference. But the success rate really took off once researchers began to build analogical reasoners based on neural network models – artificial reasoners that encode and process information the way a biological brain does. One of the most influential and ambitious of these is LISA, a neural network system created by John Hummel that embodies Holyoak and Thagard's multi-constraint theory (Hummel and Holyoak, 2005). The genius underlying this system is that it seeks out analogies by looking for similarities at each level of abstraction – surface features, simple relations, and higher-order relations. And it does it using brain-like processes.

To understand why neural networks so successfully model natural cognition, you have to understand a little about how your brain works. The basic processing units of the brain are nerve cells called neurons. *Neurons are instruments of communication*. They receive, process, and send information. This information comes to them in the form of electrochemical signals. When your retina receives light, it sends an electrical signal to the brain via the optic nerve. This electric signal is created by changing the permeability of the cell's wall so that there is an ion exchange. A given cell receives a lot of these signals from other cells. When the summated signal strength gets high enough, the cell "fires," sending its signal down a long branch called an axon. When the end of the axon is reached, a cocktail of chemicals called neurotransmitters are released that travel across the gap between that cell and another cell. This gap is called a synapse. Your brain learns by modifying synapses – making them more or less likely to act on signals from other cells.

So at bottom, thinking is an electrochemical activity spreading through a network of neurons, and the act of learning (no matter what kind) modifies those nerve cells.

Now what if you could build a machine that works like that? We have, and they are called neural networks. Neural networks are made of artificial neurons that behave like biological neurons. These artificial neurons can be programmed to represent concepts (such as "light bulb" or "bridge"), features (such as "fragile" or "narrow"), and other types of knowledge (such as "cause-of"). These neurons are organized in layers. The bottom layer receives inputs from the environment, the top layer sends outputs (such as decisions) to the environment, and the middle layers process inputs to turn them into reasonable outputs.

In the network, neurons are connected to one another with a weighted link. The weight is a measure of how strongly related the two neurons are. If we were to build a network to represent the above problems, the features that describe the light-bulb problem would all have large weights on the links that connect them. The same would be true of the army problem. But the links *between* the stories would have very weak links. This means network would understand each story but would not think they have much to do with each other.

Now suppose the network was given Duncker's x-ray problem to solve. An x-ray is a type of light energy, so "light bulb" would become activated; this means the network would be reminded of the light-bulb story. X-ray is also similar to a laser, so "laser" would also be activated, and their activation would spread through the network. All of the strongly linked neurons would become strongly active, but the weakly linked ones would be only weakly activated. This means network is "reminded" of the light-bulb story. The army story is also active, but its activation levels would be so low that they wouldn't have much impact on the network's subsequent processes. If the network were human, we would say that the army story was niggling at the periphery of consciousness, but the light-bulb story was plainly present in consciousness.

So what does this mean? It means this network has "seen" the analogy between the x-ray problem and the light-bulb story. The convergence solution from the light-bulb story will be activated and hence available for solving the X-ray problem. So we have modeled memory, consciousness, and analogical reasoning with a bunch of really simple units and processes.

Moreover, if the network knew only about the army story, it would not have made much of a connection between the two stories. There may have been mild activation of the army story, but not enough to influence the network's final decision. So, just like people, even though it had a perfectly good analog in memory that could be used to solve the x-ray problem, it would not be reminded of it.

Like brains, neural networks learn by modifying connection weights – making a neuron more or less likely to fire when it receives activation from another neuron. So suppose we were to tell the network "Hey, the light-bulb problem and the army problem are the same type of problem because you can use the convergence solution to both." This would force the network to compare the two problems. It will "reward" the links that would have gotten it to "converge" (such as "preserve surround") by increasing the connection weight, and "punish" the links that made it fail to see the similarity. Notice that the neurons that made it fail are the surface-feature neurons. The two problems have nothing in common at that level, so the connection weights would be decreased dramatically. The network now has linked the two problems in memory by abstracting over irrelevant details. If it were presented with the x-ray problem now, the higher-order relations in the problem would match the remaining representations of both analogs, and the network would "see" the similarity between all three problems.

Notice that we didn't need to write any special rules or strategies to get analogical reasoning out of this system. It is just the way neural networks work. They are similarity-driven, and similarity can work on a number of different levels of abstraction. Just like the brain, these networks are activated when incoming information matches the information in the network. Just like the brain, that activation spreads through the network. Just like the brain, whatever is strongly activated constitutes working memory. Just like the brain, the network learns by changing the strength of connections among neurons. The natural and automatic output of both systems is analogical reasoning. (Caveat: Of course, artificial systems have no self-awareness; they are not conscious. But if they were, what they are currently conscious of could be determined just by looking at which neurons are currently highly activated.)

Brains and artificial neural networks also share another extremely useful characteristic: Both learn simply through exposure to examples. In fact, both learn best that way. This means that examples are crucially important for instruction; as any good teacher knows, one good example is worth an hour of lecture. This makes a programmer's or system designer's job immensely easier. They don't need to teach an artificial neural network about a category of object or events. They just need to give it lots of examples, let it try to classify them based on similarity, and tell it when it got the classifications right or wrong. It will do the rest. Neural networks can even create their own algorithms for doing the job better.

Finally, what about that "implicit" activation that isn't strong enough to achieve consciousness? In the last chapter, we saw that insightful solutions often begin as weak activation in non-dominant areas of the brain. Neural networks readily model this type of problem solving. And they explain those "flashes of brilliance" that seem to appear out of the blue, sometimes changing entire disciplines of knowledge. Here are some examples from the history of science:

- Friedrich August Kekule von Stradonitz (1829–1896) was a German chemist who struggled to discover the structure of the chemical compound benzene – until he had a dream of whirling snakes, one of which appeared to be swallowing its tail. The structure of benzene is like a snake swallowing its tail. It is a circular structure of carbon atoms.
- Dmitri Mendeleev (1834–1907) dreamed of "a table where all the elements fell into place as required" an insight that led him to propose the periodic table of the elements where elements were listed by atomic weight and valence.
- Otto Loewi (1873–1961) believed nerve impulses were chemical, but found it difficult to design an experiment that would prove the matter. One night, he dreamed of just such an experiment. He excitedly scribbled the experiment onto a scrap of paper on his nightstand and went back to sleep. When he awoke the next morning, he found, to his horror, that he couldn't read what he'd written and he couldn't remember the dream! Fortunately, he had the same dream that night. Leaving nothing up to chance, he immediately went to his laboratory and conducted the experiment, which was successful and earned him the Nobel Prize in chemistry.

We all experience analogy-based insights daily because, as Holyoak and Thagard (1997) put it, the "analogical mind is simply the mind of a normal human being." But these amazing discoveries seem like the output of analogical minds on steroids.

Box 9.1. Some of the Best (and Funniest) Analogies from High School Students

The *Washington Post* held a contest in which high school teachers sent in the worst analogies they had encountered in grading their students' papers over the years. Here are some that are actually terrifically witty:

1. He fell for her like his heart was a mob informant and she was the East River.
2. The plan was simple, like my brother-in-law Phil. But unlike Phil, this plan just might work.
3. It came down the stairs looking very much like something no one had ever seen before.
4. The dandelion swayed in the gentle breeze like an oscillating electric fan set on medium.
5. Her lips were red and full, like tubes of blood drawn by an inattentive phlebotomist.
6. Her vocabulary was as bad as, like, whatever.
7. She grew on him like she was a colony of E. coli and he was room-temperature Canadian beef.
8. The lamp just sat there, like an inanimate object.
9. It was an American tradition, like fathers chasing kids around with power tools.
10. She was as unhappy as when someone puts your cake out in the rain, and all the sweet green icing flows down and then you lose the recipe, and on top of that you can't sing worth a damn.
11. His thoughts tumbled in his head, making and breaking alliances like underpants in a dryer without Cling Free.
12. He was as tall as a 6′3″ tree.
13. Her face was a perfect oval, like a circle that had its two sides gently compressed by a Thigh Master.
14. From the attic came an unearthly howl. The whole scene had an eerie, surreal quality, like when you're on vacation in another city and *Jeopardy* comes on at 7 p.m. instead of 7:30.
15. John and Mary had never met. They were like two hummingbirds who had also never met.
16. She had a deep, throaty, genuine laugh, like that sound a dog makes just before it throws up.

(*continued*)

17. He was as lame as a duck. Not the metaphorical lame duck, either, but a real duck that was actually lame. Maybe from stepping on a land mine or something.
18. Long separated by cruel fate, the star-crossed lovers raced across the grassy field toward each other like two freight trains, one having left Cleveland at 6:36 p.m. traveling at 55 mph, the other from Topeka at 4:19 p.m. at a speed of 35 mph.
19. The young fighter had a hungry look, the kind you get from not eating for a while.
20. The sardines were packed as tight as the coach section of a 747.

References

Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, *54*, 299–352.

Anderson, J. L. (1998). Embracing uncertainty: The interface of Bayesian statistics and cognitive psychology. *Conservation Ecology*, *2*(1): 2. Available online at http://www.consecol.org/vol2/iss1/art2/

Armstrong, D. (1983). *What is a law of nature?* Cambridge: Cambridge University Press.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, *211*, 1390–1396.

Bacon, F. (1620). *Novum Organum.* Available online at http://www.constitution.org/bacon/nov_org.htm

Ball, W. (1973). The perception of causality in the infant. Presented at the Meeting of the Society for Research in Child Development, Philadelphia, PA.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chance. *Philosophical Transactions of the Royal Society of London*, *53*(10), 370–418.

Bentham, J. (1789/2005). *Utilitarianism in principles of morals and legislation.* Adamant Media Corporation.

Borg, J. S., Hynes, C., Van Horn, J., Grafton, S., & Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: An fMRI investigation. *Journal of Cognitive Neuroscience*, *18*, 803–817.

Bowden, E. M., Jung-Beeman, M., Fleck, J., & Kounios, J. (2005). New approaches to demystifying insight. *Trends in Cognitive Science*, *9*, 322–328.

Bowers, K. S., Rehehr, G., Balthazard, C., & Parker, K. (1990). Intuition in the context of discovery. *Cognitive Psychology*, *22*, 72–110.

Breiter, H. C., Aharon, I., Kahneman, D., Dale, A., & Shizgal, P. (2001). Functional imaging of neural responses to expectancy and experience of monetary gains and losses. *Neuron*, *30*, 619–639.

Buist, D. S., Anderson, M. L., Haneuse, S. J., Sickles, E. A., Smith, R. A., et al. (2011). Influence of annual interpretive volume on screening mammography performance in the United States. *Radiology*, *259*, 72–84.

Burstein, M. (1986). Concept formation by incremental analogical reasoning and debugging. In R. Michalski et al. (Eds.), *Machine learning: An artificial intelligence approach* (Vol. 2, pp. 351–370). New York: Morgan Kaufman.

Byrne, R. W., & Russon, A. E. (1998). Learning by imitation: a hierarchical approach. *Behavioral & Brain Sciences*, *21*, 667–721.

Camerer, C. F. (2003). Behavioral studies of strategic thinking in games. *Trends in Cognitive Sciences*, *7*, 225–231.

Cartwright, N. (1980). Do the laws of physics state the facts? *Pacific Philosophical Quarterly*, *61*, 75–84.

Chan, D., & Chua, F. (1994). Suppression of valid inferences: Syntactic views, mental models, and relative salience. *Cognition*, *53*, 217–238.

Chase, V. M., Hertwig, R., & Gigerenzer, G. (1998). Visions of rationality. *Trends in Cognitive Sciences*, *2*, 206–214.

Chase, W., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*, 55–81.

Cheney, D. L., & Seyfarth, R. M. (1992). *How monkeys see the world: Inside the mind of another species.* Chicago: University of Chicago Press.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367–405.

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review*, *99*, 365–382.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*, 121–152.

Christiansen, C. L., Wang, F., Barton, M. B., Kreuter, W., Elmore, J. G., Gelfand, A. E., & Fletcher, S. W. (2000). Predicting the cumulative risk of false-positive mammograms. *Journal of the National Cancer Institute*, *92*, 1657–1666.

Cummins, D. D. (1992). Role of analogical reasoning in the induction of problem categories. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *18*, 1103–1124.

(1995). Naive theories and causal deduction. *Memory & Cognition*, *23*, 646–658.

(1997). Reply to Fairley and Manktelow's comment on "Naïve theories and causal deduction." *Memory & Cognition*, *25*, 415–416.

Cummins, D. D., Lubart, T., Alksnis, O., & Rist, R. (1991). Conditional reasoning and causation. *Memory & Cognition*, *19*, 274–282.

Dawkins, S. (1976). *The selfish gene*. Oxford: Oxford University Press.

De Martino, B., Kumaran, D., Seymour, B., & Dolan, R. J. (2006). Frames, biases, and rational decision-making in the human brain. *Science*, *313*, 684–687.

de Neys, W., Schaeken, W., & d'Ydewalle, G. (2002). Causal conditional reasoning and semantic memory retrieval: A test of the "semantic memory framework." *Memory & Cognition*, *30*, 908–920.

(2003). Inference suppression and semantic memory retrieval: Every counterexample counts. *Memory & Cognition*, *31*, 581–595.

de Neys, W., Vartanian, O., & Goel, V. (2008). Smarter than we think: When our brains detect that we are biased. *Psychological Science*, *19*, 483–489.

de Quervain, D. J., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science*, *305*, 1254–1258.

de Waal, F. (1989). Food sharing and reciprocal obligations among chimpanzees. *Journal of Human Evolution*, *18*, 433–459.

Duncker, K. (1945). On problem solving. *Psychological Monographs*, *58*, Whole No. 270.

Eckel, C. C., & Grossman, P. J. (1995). Altruism in anonymous dictator games. *Games and Economic Behavior*, *16*, 181–191.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgements under uncertainty: Heuristics and biases* (pp. 249–267). Cambridge: Cambridge University Press.

Edwards, W. (1955). Experimental measurement of utility. *Econometrica*, *23*, 346–347.

Elio, R. (1998). How to disbelieve p –> q: Resolving contradictions. *Proceedings of the Twentieth Meeting of the Cognitive Science Society*, 315–320.

Evans, J. St. B., Barston, J., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, *11*, 295–306.

Evans, J. St. B. T., & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual process theory of reasoning. *Thinking & Reasoning*, *11*, 382–389.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine. *Artificial Intelligence*, *41*, 1–63.

Fehr, E., & Fischbacher, U. (2004a). Social norms and human cooperation. *TRENDS in Cognitive Sciences*, *8*, 185–190.

(2004b). Third-party punishment and social norms. *Evolution & Human Behavior*, *25*, 63–87.

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, *90*, 980–994.

Fehr, E., & Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature*, *422*, 137–140.

Ferguson, R. W. (1994). MAGI: Analogy-based encoding using symmetry and regularity. In *Proceedings of the 16th Annual Conference of Cognitive Science Society* (pp. 283–288). Mahwah, NJ: Lawrence Erlbaum Associates.

Fiddick, L., & Cummins, D. D. (2007). Are perceptions of fairness relationship specific? The case of noblesse oblige. *Quarter Journal of Experimental Psychology*, *60*, 6–31.

Financial Crisis Inquiry Commission (2011). *The Financial Crisis Inquiry Report, Authorized Edition: Final Report of the National Commission on the Causes of the Financial and Economic Crisis in the United States*. Jackson, TN: Public Affairs.

Foot, P. (1978). *The problem of abortion and the doctrine of the double effect in virtues and vices.* Oxford: Basil Blackwell.

Fugelsang, J., & Dunbar, K. (2005). Brain-based mechanisms underlying complex causal thinking. *Neuropsychologia*, *43*, 1204–1213.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*, 155–170.

Gick, M. L., and Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*, 306–355.

Gigerenzer, G. (1994). Why the distinction between single-event probabilities and frequencies is relevant for psychology (and vice versa). In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 129–161). New York: Wiley.

(2002). *Calculated risks: How to know when numbers deceive you*. New York: Simon & Schuster.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2008). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8*, 53–96.

Gigerenzer, G., Hell, W., & Blank, H. (1988). Presentation and content: The use of base rates as a continuous variable. *Journal of Experimental Psychology: Human Perception & Performance*, *14*, 513–525.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*, 684–704.

Gigerenzer, G., Todd, P., & the ABC group (2000). *Simple heuristics that make us smart*. Oxford: Oxford University Press.

Gobet, F., & Simon, H. A. (1996). The roles of recognition processes and look-ahead search in time-constrained expert problem solving: Evidence from grand-master-level chess. *Psychological Science*, *7*, 52–55.

Goel, V., & Dolan, R. J. (2003). Explaining modulation of reasoning by belief. *Cognition*, *87*, B11–B22.

Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences*, *11*, 322–323.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*, 389–400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*, 2105–2108.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Strength and structure in causal induction. *Cognitive Psychology*, *51*, 334–384.

Haidt, J. (2007). The new synthesis in moral psychology. *Science*, *316*, 998–1002.

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, *20*, 98–116.

Haidt, J., & Joseph, C. (2008). The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. In P. Carruthers, S. Laurence & S. Stich (Eds.), *The innate mind Volume 3: Foundations and the future* (pp. 367–391). New York: Oxford University Press.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation in preverbal infants. *Nature*, *450*, 557–559.

Harris, L. (2007). *The suicide of reason: Radical Islam's threat to the West.* New York: Basic Books.

Heekeren, H. R., Wartenburger, I., Schmidt, H., Schwintowski, H., & Villringer, A. (2003). An fMRI study of simple ethical decision-making. *Neuroreport*, *14*, 1215–1219.

Hertwig, R., & Gigerenzer, G. (1999). The "conjunction fallacy" revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making*, *12*, 275–305.

Hoffman, E., McCabe, K., Shachat, K., & Smith, V. (1994). Preferences, property rights, and anonymity in bargaining games. *Games & Economic Behavior*, *7*, 346–380.

Hoffman, E., McCabe, K., & Smith, V. (1996). Social distance and other-regarding behavior in dictator games. *American Economic Review*, *86*, 653–660.

Hoffman, E., & Spitzer, M. (1985). Entitlements, rights, and fairness: An experimental examination of subjects' concepts of distributive justice. *Journal of Legal Studies*, *15*, 254–297.

Hofman, W., & Baumert, A. (2010). Immediate affect as a basis for intuitive moral judgement: An adaptation of the affect misattribution procedure. *Cognition & Emotion*, *24*, 522–535.

Hofstadter, D. (2009). Analogy as the core of cognition. In D. Gentner, K. Holyoak & B. Kokinov (Eds.), *The Analogical mind: Perspectives from cognitive science* (pp. 499–538). Cambridge, MA: The MIT Press.

Holroyd, C.B., Larsen, J.T., & Cohen, J.D. (2004). Context dependence of the event-related brain potential associated with reward and punishment. *Psychophysiology*, *41*, 245–253.

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition*, *15*, 332–340.

Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science*, *13*, 295–355.

(1997). The analogical mind. *American Psychologist*, *52*, 35–44.

Hume, D. (1740/1967). *A treatise of human nature*. Oxford: Oxford University Press.

(1748/2010). *An enquiry concerning human understanding.* New York: General Books.

(1751/1998). *An enquiry concerning the principles of morals*. Oxford: Oxford University Press.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*, 427–466.

Hummel, J. E., & Holyoak, K. J. (2005). Relational reasoning in a neurally plausible cognitive architecture: An overview of the LISA project. *Current Directions inCognitive Science*, *14*, 153–157.

Inglis, M., & Simpson, A. (2007). Belief bias and the study of mathematics. In D. Pitta-Pantazi & G. Philippou (Eds.), *Proceedings of the Fifth Congress of the European Society for Research in Mathematics Education* (pp. 2310–2319). Larnaca, Cyprus.

Janveau-Brennan, G., & Markovits, H. (1999). The development of reasoning with causal conditionals. *Developmental Psychology*, *35*, 904–911.

Juslin, P., Winman, A. & Olsson, H. (2000). Naïve empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psychological Review*, *107*, 384–396.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*, 697–720.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*, 237–251.

(1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–291.

Kant, I. (1781/2008). *Critique of pure reason.* New York: Penguin Classics.

(1783/1911). *Prolegomena to any future metaphysics. Akademie* edition, vol. IV, Berlin.

(1785/1989). *The foundations of the metaphysics of morals*. Upper Saddle River, NJ: Prentice-Hall.

(1787/1997). *The critique of practical reason*. Cambridge: Cambridge University Press.

Kaufman, J., & Zigler, E. F. (1988). Do abused children become abusive parents? *Annual Progress in Child Psychiatry & Child Development*, *29*, 591–600.

Keller, G. (2011). *Statistics for management and economics.* Chula Vista, CA: Southwestern College Publishing.

Kelley, H. H., & Stahelski, A. J. (1970). The inference of intention from moves in the Prisoner's Dilemma Game. *Journal of experimental social psychology*, *6*, 401–419.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: Reputation and trust in a two-person economic exchange. *Science*, *308*, 78–83.

Kirsch, I., Deacon, B. J., Huedo-Medina, T. B., Scoboria, A., Moore, T. J., & Johnson, B. T. (2008). Initial severity and antidepressant benefits: A meta-analysis of data submitted to the Food and Drug Administration. *PLoS Medicine*, 5(2).

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*, 211–228.

Knutson, B., & Bossaerts, P. (2007). Neural antecedents of financial decisions. *Journal of Neuroscience*, *27*, 8174–8177.

Knutson, B., Taylor, J., Kaufman, M., Peterson, R., & Glover, G. (2005). Distributed neural representation of expected value. *Journal of Neuroscience*, *25*, 4806–4812.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgments. *Nature*, *446*, 908–911.

Kosfeld, M., Heinrichs, M., Zaks, P. J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature*, *435*, 673–676.

Kotovsky, L., & Baillargeon, R. (2000). Reasoning about collision events involving inert objects in 7.5-month-old infants. *Developmental Science*, *3*, 344–359.

Kuehne, S. E., Forbus, K. D., Gentner, D., & Quinn, B. (2000). SEQL: Category learning as progressive abstraction using structure mapping. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.

Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Kurtz, K. J., & Loewenstein, J. (2007). Converging on a new role for analogy in problem solving and retrieval: When two problems are better than one. *Memory & Cognition*, *35*, 334–341.

Lakshminarayanan, V. R., Chen, M. K., & Santos, L. (2011). The evolution of decision-making under risk: Framing effects in monkey risk preferences. *Journal of Experimental Social Psychology*, *47*, 689–693.

Lewis, D. (1973). Causation. *Journal of Philosophy*, *70*, 556–567.

(1979). Counterfactual dependence and time's arrow. *Nous*, *13*, 455–476.

(2000). Causation as influence (abridged version). *Journal of Philosophy*, *97*, 182–197.

(2004). Void and object. In J. Collins, N. Hall & L. A. Paul (Eds.), *Causation and counterfactuals* (pp. 277–290). Cambridge, MA: MIT Press.

Lichtenstein, S., Fischhoff, B., & Phillips, L. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–344). Cambridge/New York: Cambridge University Press.

Lombrozo, T. (2009). The role of moral commitments in moral judgment. *Cognitive Science*, *33*, 273–286.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*, 2098–2109.

Luchins, A. (1942). *Mechanization in problem solving. Psychological Monographs 34*. Washington, DC: American Psychological Association.

Luchins, A., & Luchins, E. H. (1959). Rigidity of behavior – a variational approach to the effect of einstellung. Eugene: University of Oregon Books.

Luo, J., Yuan, J., Qiu, J., Zhang, Q., Zhong, J., & Huai, Z. (2008). Neural correlates of the belief-bias effect in syllogistic reasoning: An event-related potential study. *NeuroReport*, *19*, 1075–1080.

Mackie, J. L. (1974). *The cement of the universe: A study in causation.* Oxford: Clarendon.

Martin, J. A., & Elmer, E. (1992). Battered children grown up: A follow-up study of individuals severely maltreated as children. *Child Abuse & Neglect*, *16*, 75–88.

Maynard Smith, J. (1972). *On evolution*. Edinburgh: Edinburgh University Press.

McCarthy, J. (1980). Circumscription – a form of non-monotonic reasoning. *Artificial Intelligence*, *13*, 27–39.

Metcalfe, J. (1986). Feeling of knowing in memory and problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*, 288–294.

Metcalfe, J., & Wiebe, D. (1987). Intuition in insight and non-insight problem solving. *Memory & Cognition*, *15*, 238–246.

Michotte, A. (1963). *The perception of causality.* New York: Basic Books.

Mill, J. S. (1843/1963). *System of logic, ratiocinative and inductive.* Reprinted in J. M. Robson (Ed.), *Collected Works of John Stuart Mill.* Toronto: University of Toronto Press.

(1859/2011). *On liberty.* Hollywood, FL: Simon Brown.

(1861/2007). *Utilitarianism.* Mineola, NY: Dover.

(1869/2011). *Subjection of women.* Clippesby: Croft Classics.

Moll, E., & de Oliveira-Souza (2001). Frontopolar and anterior temporal cortex activation in a moral judgment task: Preliminary function MRI results in normal subjects. *Arq Neuropsiquiatr*, *59*, 657–664.

Morgenstern, O., & von Neumann, J. (1944). *Theory of games and economic behavior.* Princeton, NJ: Princeton University Press.

Muentener, P., & Carey, S. (2010). Infants' causal representations of state change events. *Cognitive Psychology*, *61*, 63–86.

Nash, J. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, *36*(1), 48–49.

National Cancer Institute Surveillance, Epidemiology, and End Results (SEER). Retrieved from http://seer.cancer.gov/statfacts/html/breast.html

Newborn, M. (1997). *Kasparov versus Deep Blue: Computer chess comes of age*. New York: Springer.

Newell, A., & Simon, H. (1963). GPS: A program that simulates human thought. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought* (pp. 279–293). New York: McGraw-Hill.

Oaksford, M., & Chater, N. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Science*, *3*, 57–65.

Oaksford, M., Chater, N., Grainger, B., & Larkin, J. (1997). Optimal data selection in the reduced array selection task (RAST). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 441–458.

O'Doherty J., Deichmann R., Critchley H. D., & Dolan R. J. (2002). Neural responses during anticipation of a primary taste reward. *Neuron*, *33*, 815–826.

Olsen, V. (2004). Man of action. *Smithsonian Magazine*, September. Retrieved from http://www.smithsonian.com

O'Rourke, P. J. (1999). *Eat the rich.* London: Atlantic Monthly Press.

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* Cambridge: Cambridge University Press.

(2009). Causal inference in statistics: An overview. *Statistics Surveys*, *3*, 96–146.

Peirce, C. (1877). The fixation of belief. *Popular Science Monthly*, 1–15.

Pollock, J. (1987). Defeasible reasoning. *Cognitive Science*, *11*, 481–518.

Polya, G. (1945). *How to solve it.* Princeton, NJ: Princeton University Press.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *American. Economics Review*, *83*, 1281–1302.

Reingold, E. M., Charness, N., Pomplun, M., & Stampe, D. M. (2001). Visual span in expert chess players: Evidence from eye movements. *Psychological Science*, *12*, 48–55.

Rilling, J., Gutman, D., Zeh, T., Pagnoni, G., Berns, G., & Kilts, C. (2002). A neural basis for social cooperation. *Neuron*, *35*, 395–405.

Ross, B. H. (1984). Remindings and their effects in learning a cognitive skill. *Cognitive Psychology*, *16*, 371–416.

Russell, B. (1959). *My philosophical development.* London/New York: George Allen and Unwin/Simon and Schuster.

Sanfey, A. G. (2007). Social decision-making: Insights from game theory and neuroscience. *Science*, *318*, 598–602.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the Ultimatum Game. *Science*, *300*, 1755–1758.

Schnall, S., Benton, J., & Harvey, S. (2008a). With a clean conscience: Cleanliness reduces the severity of moral judgments. *Psychological Science*, *19*, 1219–1222.

Schnall, S., Haidt, J., Clore, G., & Jordan, A. (2008b). Disgust as embodied moral judgment. *Personality & Social Psychology Bulletin*, *34*, 1096–1109.

Schwitzgebel, E., & Cushman, F. (in press). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind & Language.*

Silveira, J. (1971). Incubation: The effect of interruption timing and length on problem solution and quality of problem processing. Reported in J. R. Anderson (1985). *Cognitive psychology and its implications.* Basingstoke: W. H. Freeman.

Sloman, S. A., Over, D., Slovak, L., & Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organizational Behavior & Human Decision Processes*, *91*, 296–309.

Smith, A. McCall (2007). *Love over Scotland.* New York: Anchor Books.

Sober, E., & Sloan-Wilson, D. (1999). *Unto others: The evolution and psychology of unselfish behavior.* Cambridge, MA: Harvard University Press.

Sohn, E. (2001). Stopping time in its tracks. *U.S. News & World Report*, July 9. (Also electronically reprinted on www.usnews.com)

Spiro, M. E. (1996). Postmodernist anthropology, subjectivity, and science: A modernist critique. In *Comparative studies in society and history* (Vol. 5, pp. 759–780). Ann Arbor: University of Michigan Press.

Thomson, J. J. (1985). The trolley problem. *Yale Law Journal*, *94*, 1395–1415.

Trivers, R. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, *46*, 35–57.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453–458.

(1982). Judgments of and by representativeness. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 84–98). New York: Cambridge University Press.

(1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.

U.S. Securities and Exchange Commission and the Commodity Futures Trading Commission (2010). Findings regarding the market events of May 6, 2010. September 30.

van Dijk, E., & Vermunt, R. (2000). Strategy and fairness in social decision making: Sometimes it pays to be powerless. *Journal of Experimental Social Psychology*, *36*, 1–25.

Vershueren, N., Schaeken, W., de Neys, W., & d'Ydewalle, G. (2004). The difference between generating counterexamples and using them during reasoning. *Quarterly Journal of Experimental Psychology*, *57*, 1285–1308.

U.S. Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*, 129–140.

(1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*, 273–281.

Weg, E., & Smith, V. (1993). On the failure to induce meager offers in ultimatum games. *Journal of Economic Psychology*, *14*, 17–32.

Westin, D., Blagov, P. S., Harenski, K., Kilts, C., & Hamann, S. (2006). Neural bases of motivated reasoning: An fMRI study of emotional constraints on partisan political judgment in the 2004 U.S. presidential election. *Journal of Cognitive Neuroscience*, *18*, 1947–1958.

White, P. A. (1995). Use of prior beliefs in the assignment of causal roles: causal powers versus regularity-based accounts. *Memory & Cognition*, *23*, 43–54.

(2000). Causal judgment from contingency information: The interpretation of factors common to all instances. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 1083–1102.

Whitehead, A. N., & Russell, B. (1910, 1912, 1913). *Principia mathematica* (3 vols). Cambridge: Cambridge University Press.

Wilkinson, G. S. (1984). Reciprocal food sharing in the vampire bat. *Nature*, *308*, 181–184.

Williams, A. M., Davids, K., Burwitz, L., & Williams, J. G. (1992). Perception and action in sport. *Journal of Human Movement Studies*, *22*, 147–204.

(1994). Visual search strategies in experienced and inexperienced soccer players. *Research Quarterly for Exercise and Sport*, *65*, 127–135.

Wilson, W. A., & Kuhn, C. M. (2005). How addiction hijacks our reward system. *Cerebrum*, *7*, 53–66.

Woodruff, G., & Premack, D. (1979). Intentional communication in the chimpanzee: The development of deception. *Cognition*, *7*, 333–362.

Young, L., Camprodon, J. A., Hauser, M., Pasual-Leones, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences*, *107*, 6753–6758.

Zacks, R., & Hasher, L. (2002). Frequency processing: A twenty-five-year perspective. In P. Sedlmeier & T. Betsch (Eds.), *Frequency processing and cognition* (pp. 21–36). New York: Oxford University Press.

Zak, P. J., Stanton, A. A., & Ahmadi, S. (2007). Oxytocin increases generosity in humans. *PLoS ONE 2*(11): e1128. doi:10.1371/journal.pone.0001128.

Index

addiction, 54, 55.
See also neural reward circuitry
algorithms, 51, 140–146, 152, 155, 160, 171, 182
Alhacen, 121
Ali, Ayaan Hirsi, 3
altruism, 15, 16, 19, 24, 27, 28
pure, 20
reciprocal, 27–29, 78
analogical reasoning, 84, 168–182
and artificial intelligence, 178–182
definition, 170
analogy. *See* analogical reasoning
Aristotle, 96, 97, 119, 121, 135
Armstrong, David, 109
artificial intelligence, 94, 136, 152–155, 178–182.
See also neural networks
and expert systems, 154–155
Axelrod, Robert, 13, 14, 26

Bacon, Sir Francis, 118, 121, 122
Bayes Rule, 35, 36, 37, 38
Bayes, Thomas, 35. *See also* Bayes Rule; Bayesian reasoning
Bayesian reasoning, 6, 35–44, 126, 133.
See also probabilistic reasoning
behaviorism, 136, 137, 152, 159
belief bias. *See* biases
Bentham, Jeremy, 62, 70–72
biases. *See also* framing effect
belief bias, 92–94, 106
confirmation bias, 116, 117, 120, 124
loss aversion bias, 47, 48
overconfidence bias, 50–51, 58
prototype, 46, 55, 56
Boole, George, 97

cancer screening, 31, 33–42. *See also* HIV screening
Cartwright, Nancy, 109
causal logic. *See* logic, modal
causal necessity and causal sufficiency, 90, 103, 106–109, 110
causal reasoning, 99–113
and disablers, 96, 109–110
in infants, 111–113
models of, 101, 103–106
Chase, William, 156
Chater, Nick, 125, 126
Cheng, Patricia, 101, 102, 103, 110
chess, 7, 8, 155–157
child abuse, 109, 110
confirmation bias. *See* biases
conjunction fallacy, 45
cooperation, 12–19. *See also* altruism, reciprocal
evolution of, 26–30
tit-for-tat, 14
Cummins, Denise, 21, 22, 95, 96, 109, 110, 177
Cummins, Robert, 3

Dawkins, Richard, 26
deductive. *See also* logic
reasoning, 91–94
validity, 82–84, 96, 97
defeasible reasoning, 94–96
deontology, 65–70
depression, 124–128, 131–132. *See also* SSRI
developmental psychology, 26, 111, 112, 169,
See also infant cognition
dual-process theory 49, 50, 56, 73, 74, 76, 93
Duncker, Karl, 139, 140, 150, 152, 158, 165, 175, 177, 180. *See also* x-ray problem
Durkheim, Emile, 77

Einstellung effect, 163, 171. *See also* framing effect
emotion, 2, 24, 49, 55, 93, 120. *See also* moral decision making and emotion
equilibrium, 7, 8, 10, 11, 12. *See also* game theory
  Nash, 8, 9, 10, 11, 12, 15
evolutionarily stable strategy, 27, 29
experimental economics, 15–23

Foot, Philippa, 59
framing effect, 47, 55, 163
Frege, Gottlob, 98
frontal cortex, 23, 53, 93

game theory, 5–29
games. *See also* Prisoner's Dilemma
  cooperative and non-cooperative, 7, 8
  Dictator, 19–21, 22, 24
  public goods, 16, 17, 18, 19, 24, 27
  Trustee, 18–19, 24
  Ultimatum, 19–21, 22, 24, 25
Gentner, Dedre, 171, 172, 178. *See also* Structure Mapping Theory
Gigerenzer, Gerd, 14, 31, 39, 40, 41, 42, 46, 51. *See also* heuristics
Gödel, Kurt, 98
Greene, Joshua, 72, 73, 74, 75

Ha, Young-Won, 123
Haidt, Jonathan, 73, 74, 77, 78
Hamilton, W.D., 14, 26, 28. *See also* kin selection
Harris, Lee, 2
heuristics, 46, 49, 51–53, 56, 123, 146–148, 149, 150, 160
  means-end analysis, 150–151, 152
HIV screening, 43
Hofstadter, Douglas, 169
Holyoak, Keith, 169, 172, 175, 177, 179, 182
Hume, David, 62–65, 66, 68, 71, 73, 75, 99–100, 101, 104, 107, 113, 122
hypothesis testing, 1, 115–134
  and falsification, 122, 124, 127, 137
hypothetico-deductive method, 121, 127, 137

inductive reasoning, 83, 84
infant cognition, 25, 111–113
insight, 158–165

Jefferson, Thomas, 61

Kahneman, Daniel, 43–45, 46, 47, 49, 53, 74
Kant, Immanuel, 62, 65–70, 71, 72, 76, 100, 113
Kasparov, Garry, 156
kin selection, 26, 28, 30. *See also* altruism
Klayman, Joshua, 123
Köhler, Wolfgang, 158, 159
Kuhn, Thomas, 134, 135. *See also* paradigm shifts

Lewis, David, 107, 108
Loewi, Otto, 182
logic, 80–98
  deontic, 90
  first order, 88, 89, 94, 152
  modal, 89–91, 107
  truth functional, 85–91, 125
loss aversion bias. *See* biases
Luchins, Abraham, 163, 164

Mackie, John, 99, 108
mathematics, 84–85, 98, 177–178
Mendeleev, Dmitri, 182
Metcalfe, Janet, 160, 165
Michotte, Albert, 111, 112, 113
Mill, John Stuart, 62, 70–72, 108. *See also* utilitarianism
moral decision making, 72–79
  and emotion, 62–63, 73–77, 78
  cognitive foundations, 77–78
  and political orientation, 78, 79
moral dilemmas, 60, 61, 72, 75–76. *See also* trolley problem
moral philosophy, 1, 62–72. *See also* deontology; utilitarianism
Morgenstern, Oskar, 6

Nash, John, 8, 10, 11, *See also* equilibrium, Nash
neural networks 179–182
neural reward circuitry, 23–25, 53–55, 120. *See also* addiction
neuroscience, 23–25, 53–56, 74–77, 93–94, 105–106, 119–120, 161–162
Newell, Alan, 152
norms, 16, 18, 19, 20, 21, 33, 77
Novick, Laura, 101, 102, 103

Oaksford, Mike, 125, 126
optimal data selection. *See* Bayesian reasoning
overconfidence. *See* biases
oxytocin, 24

paradigm shifts, 134–137
Peirce, Charles, 121
Polya, George, 139, 140
Popper, Karl, 122, 126
prefrontal cortex, 54, 55, 56, 75, 94, 120, 191

primate cognition, 27, 48, 49, 151, 159, 191
Prisoner's Dilemma, 12–19, 23, 24, 29
probabilistic reasoning, 23, 31–47, 53–56, 101–103, 130, 131, 132, 133–134, 155. *See also* Bayesian reasoning
  and frequency format. *See also* framing effect 40, 46
problem-solving, 138–167. *See also* insight
  expert, 155–158, 177
production systems. *See* artificial intelligence
propositions, 81–83, 86–88, 89, 96, 97
Prospect Theory, 49, 55
prototype. *See* biases

rational choice theory, 31–33, 53, 60
recursion. *See also* algorithms
  and language, 144, 145
Russell, Bertrand, 98

Santos, Laurie, 48, 49
selective serotonin reuptake inhibitor. *See* SSRI
Simon, Herbert, 152, 156
Smith, John Maynard, 27. *See also* evolutionarily stable strategy
social status, 19–23
SSRI. *See also* depression 127–128, 131–132
Structure Mapping Theory, 171, 172, 178. *See also* analogical reasoning

Thagard, Paul, 169, 172, 179, 182
Thomson, Judith Jarvis, 59
Tower of Hanoi, 142, 143, 152
Trivers, Robert, 27, 29, 30. *See also* altruism, reciprocal
trolley problem, 59–61, 63, 68, 69, 72
Tversky, Amos, 43, 44, 45, 46, 47, 49

utilitarianism, 70–72
utility, 6, 7, 32, 33, 35, 53, 55, 63, 70, 71, . *See also* rational choice theory; neural reward circuitry

von Neumann, John 6
von Prantl, Karl, 97
von Stradonitz, Friedrich August Kekule, 182

Wason, Peter, 115–117, 123–124, 126. *See also* hypothesis testing
water jug problems, 163
Whitehead, Alfred North, 98

x-ray problem, 150, 165, 175, 177, 180, 181