The Behavioral Foundations of Public Policy

The Behavioral Foundations of Public Policy

EDITED BY ELDAR SHAFIR

Princeton University Press
Princeton and Oxford

Published by Princeton University Press, 41 William Street,
Princeton, New Jersey 08540

In the United Kingdom: Princeton University Press, 6 Oxford Street,
Woodstock, Oxfordshire OX20 1TW

press.princeton.edu

Library of Congress Cataloging-in-Publication Data

The behavioral foundations of public policy / edited by Eldar Shafir.
p. cm.
Includes index.
ISBN 978-0-691-13756-8 (hbk. : alk. paper) 1. Social planning—Psychological aspects. 2. Political planning—Psychological aspects. 3. Policy science—Psychological aspects. I. Shafir, Eldar.
HN28.B44 2013
303.3—dc23

2012032553

British Library Cataloging-in-Publication Data is available
This book has been composed in ITC Galliard
Printed on acid-free paper. ∞
Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

Contents

Foreword vii
Daniel Kahneman

List of Contributors xi

Acknowledgments xvii

Introduction 1
Eldar Shafir

PART 1. PREJUDICE AND DISCRIMINATION

Chapter 1. The Nature of Implicit Prejudice: Implications for Personal and Public Policy 13
Curtis D. Hardin
Mahzarin R. Banaji

Chapter 2. Biases in Interracial Interactions: Implications for Social Policy 32
J. Nicole Shelton
Jennifer A. Richeson
John F. Dovidio

Chapter 3. Policy Implications of Unexamined Discrimination: Gender Bias in Employment as a Case Study 52
Susan T. Fiske
Linda H. Krieger

PART 2. SOCIAL INTERACTIONS

Chapter 4. The Psychology of Cooperation: Implications for Public Policy 77
Tom Tyler

Chapter 5. Rethinking Why People Vote: Voting as Dynamic Social Expression 91
Todd Rogers
Craig R. Fox
Alan S. Gerber

Chapter 6. Perspectives on Disagreement and Dispute Resolution: Lessons from the Lab and the Real World 108
Lee Ross

Chapter 7. Psychic Numbing and Mass Atrocity 126
Paul Slovic
David Zionts
Andrew K. Woods
Ryan Goodman
Derek Jinks

PART 3. THE JUSTICE SYSTEM

Chapter 8. Eyewitness Identification and the Legal System 145
Nancy K. Steblay
Elizabeth F. Loftus

Chapter 9. False Convictions 163
Phoebe Ellsworth
Sam Gross

Chapter 10. Behavioral Issues of Punishment, Retribution, and Deterrence 181
John M. Darley
Adam L. Alter

PART 4. BIAS AND COMPETENCE

Chapter 11. Claims and Denials of Bias and Their Implications for Policy 195
Emily Pronin
Kathleen Schmidt

Chapter 12. Questions of Competence: The Duty to Inform and the Limits to Choice 217
Baruch Fischhoff
Sara L. Eggers

Chapter 13. If Misfearing Is the Problem, Is Cost-Benefit Analysis the Solution? 231
Cass R. Sunstein

PART 5. BEHAVIORAL ECONOMICS AND FINANCE

Chapter 14. Choice Architecture and Retirement Saving Plans 245
Shlomo Benartzi
Ehud Peleg
Richard H. Thaler

Chapter 15. Behavioral Economics Analysis of Employment Law 264
Christine Jolls

Chapter 16. Decision Making and Policy in Contexts of Poverty 281
Sendhil Mullainathan
Eldar Shafir

PART 6. BEHAVIOR CHANGE

Chapter 17. Psychological Levers of Behavior Change 301
Dale T. Miller
Deborah A. Prentice

Chapter 18. Turning Mindless Eating into Healthy Eating 310
Brian Wansink

Chapter 19. A Social Psychological Approach to Educational Intervention 329
Julio Garcia
Geoffrey L. Cohen

PART 7. IMPROVING DECISIONS

Chapter 20. Beyond Comprehension: Figuring Out Whether Decision Aids Improve People's Decisions 351
Peter Ubel

Chapter 21. Using Decision Errors to Help People Help Themselves 361
George Loewenstein
Leslie John
Kevin G. Volpp

Chapter 22. Doing the Right Thing Willingly: Using the Insights of Behavioral Decision Research for Better Environmental Decisions 380
Elke U. Weber

Chapter 23. Overcoming Decision Biases to Reduce Losses from Natural Catastrophes 398
Howard Kunreuther
Robert Meyer
Erwann Michel-Kerjan

PART 8. DECISION CONTEXTS

Chapter 24. Decisions by Default 417
Eric J. Johnson
Daniel G. Goldstein

Chapter 25. Choice Architecture 428
Richard H. Thaler
Cass R. Sunstein
John P. Balz

Chapter 26. Behaviorally Informed Regulation 440
Michael S. Barr
Sendhil Mullainathan
Eldar Shafir

PART 9. COMMENTARIES

Chapter 27. Psychology and Economic Policy 465
William J. Congdon

Chapter 28. Behavioral Decision Science Applied to Health-Care Policy 475
Donald A. Redelmeier

Chapter 29. *Quis custodiet ipsos custodes?* Debiasing the Policy Makers Themselves 481
Paul Brest

Chapter 30. Paternalism, Manipulation, Freedom, and the Good 494
Judith Lichtenberg

Index 499

Foreword

DANIEL KAHNEMAN

There are no established churches in the Woodrow Wilson School of Public and International Affairs at Princeton, but there have always been established disciplines. Originally there were two: economics and politics (elsewhere known as political science). In 1999, psychology was formally introduced as the third discipline, and granted the intimidating responsibility for a semester-long compulsory class to all students working toward the degree of master of public affairs. We[1] had to find answers to some difficult questions: What does psychology have to offer to students who prepare for a career of public service? What gaps existed in our students' training that we should fill? What biases in their training should we aim to correct?

The question about biases was the easiest to answer. We observed that the students in the master's program offered by the School were exposed to a steady diet of economics courses that invoked the standard assumption of agents who are invariably rational, driven by self-interest, and motivated by tangible incentives. In the eyes of a psychologist, these propositions are not viable even as a crude approximation. The tension between psychology and the assumptions of economic theory provided a natural focus for the course we designed. Accordingly, our course emphasized errors of judgment, oddities of choice, the power of framing effects, and the intense and universal concern of people with their social group and their standing within it. We wanted our students to know that the assumptions of the rational agent model, although adequate for predicting the outcomes in many markets, are not at all adequate for predicting how individuals will actually behave in most situations. The policy-relevant situations we explored extended beyond purely economic circumstances, to issues ranging from voting and negotiations, to health behaviors, labor relations, education, and the law.

So why focus on economics in a course on psychology and policy, or in the foreword of a book about that subject? Like it or not, it is a fact of life that economics is the only social science that is generally recognized as relevant and useful by policy makers. Given their monopoly, economists have become gatekeepers, and their analyses and conclusions have enormous weight even in domains in which they do not seem to have any particular comparative advantage, such as health care and education. An obvious asymmetry in the distribution of competence contributes to the elevated status of economics: there are important policy questions that only economists are qualified to answer, but hardly any data of other social sciences that they cannot evaluate. In particular, economists have more statistical tools at their disposal than most other social scientists do. Even more important, they are native to the universal language of policy, which is money. Finally, their reputation for hard-headed objectivity gives them a significant credibility advantage over more tender-hearted practitioners of the social sciences, whom I have heard casually dismissed as "social workers."

We considered our Princeton policy students as future policy makers, who would be exposed to economic approaches to all fields of social policy. Our intent was to sensitize them to the potential pitfalls of basing policy on the standard assumptions of the rational agent model. We also mentioned to them that a growing minority of economists—behavioral economists—were engaged in attempts to develop an economic science that is based on more realistic psychological assumptions. Behavioral economics was at the time clearly defined as a distinctive approach to economics, with no particular applications to policy.

The landscape changed radically during the first decade of the new century. Behavioral economists began to address the world at large, and the boundary between behavioral economics and applied social psychology blurred, creating a new set of problems and opportunities for psychologists interested in policy. In 2001 Richard Thaler and Shlomo Benartzi reported on the success of their now famous Save More Tomorrow method for increasing workers' willingness to save from their salary. They identified three psychological obstacles to saving: loss aversion, hyperbolic discounting, and status quo bias. Save More Tomorrow was an offer to workers that bypassed these obstacles, leading them to save more. The same year, Bridget Madrian and Dennis Shea published a paper showing that an even simpler procedure—merely

changing the default—can help increase enrollments in savings plans. Now, a decade later, automatic enrollment and automatic escalation (a generic form of Save More Tomorrow) are affecting the lives and savings decisions of millions of people around the world.

A social psychologist will recognize both these strategies as brilliant reinventions of the classic Lewinian proposal for inducing behavioral change, which favors reducing the "restraining forces" over increasing the "driving forces." To follow the Lewinian approach one begins by asking "why don't people already do what I wish they would do?" This question evokes a list of restraining forces, which the agent of change then works to reduce or eliminate. The idea is transparently correct when you are exposed to it, but it is also deeply counterintuitive. The standard tools that most of us use to change others' behavior are arguments, promises, and threats. It is much less natural to look for ways of making it easier for the other person to do the right thing. Thaler and Benartzi developed a procedure that made it easy for the worker to commit to a higher saving rate in the future, which would start automatically at an auspicious time (upon receiving a salary raise). Ending the commitment, in contrast, would require a deliberate decision and a modest effort.

In subsequent articles and in their international best seller, *Nudge*, Thaler and Cass Sunstein described an approach to policy that they called "libertarian paternalism." The central idea is that it is legitimate for institutions of society to consider the best interests of individuals in structuring the choices that these individuals make—for example, about retirement saving. The goal is to make it easy and natural for casual decision makers to make sensible choices, while ensuring their complete freedom to choose as they will. This was read by all as a manifesto of the approach of behavioral economics to policy. It is founded on the ideas that the rational agent model is unrealistic, that many decisions are made with little thought, and that it is appropriate to create a "choice architecture" that reduces the incidence of foolish decisions without reducing freedom.

We have known for a long time that the role of economics in formulating policy has significant consequences. During the heyday of the rational agent model, policies were sometimes formulated that assumed rationality as a psychological fact. For example, the assumption that criminals are rational agents implies that they can be deterred by the expected disutility of being caught and punished. The probability of being caught and the severity of punishment have equivalent weights in this model, but not in reality: empirical research suggests that increasing the probability of punishment is far more effective in deterring crime than a corresponding increase of severity. In other situations, the rational agent model implies that agents need no protection against their own bad choices: choices freely made by rational agents deserve complete respect. To the surprise of most noneconomists, complete respect is often extended to awful choices, such as those that lead to addiction to noxious substances, or to lives of destitution after retirement. Because psychologists are not trained to assume that humans are rational, they are likely to find this position unattractive and even bizarre—but they recognize the risk that paternalism poses to the ideal of liberty. *Nudge* showed a way out of this dilemma: simple procedures that tend to bias people toward sensible and socially desirable choices without in any way abridging their freedom.

Nudge relied on psychology to highlight another objective that would be pointless if humans were fully rational in the role of consumers. Everyone recognizes that consumers need protection against predatory behavior, and there are many laws that are designed to provide such protection. However, the authors of *Nudge* documented many ways in which firms may take advantage of the psychological limitations of lazy and boundedly rational consumers. The book, along with work by several other researchers, showed how simple regulations can constrain predatory (though not illegal) behaviors, such as formulating truthful contracts in impenetrable language and printing them in painfully small print.

The publication of *Nudge* was immediately recognized as an important event. Sunstein became Director of the Office of Information and Regulatory Affairs (aka the "regulation czar") under President Obama, and Thaler became an advisor to a Behavioral Insight Team (colloquially known as the "Nudge Unit") established by the coalition government led by David Cameron in the UK. Other nudge units are popping up elsewhere around the world with the goal of establishing policies to help people make decisions that serve their best interest, and to protect them from exploitation in the market. The success of this enterprise can be counted as one of the major achievements of applied behavioral science in general, and of applied social and cognitive psychology in particular.

Unfortunately, because the two authors of *Nudge* were an economist and a jurist, respectively the intellectual leaders of behavioral economics and of behavioral law and economics, not only the ideas they produced themselves but also many of the contributions of cognitive and social psychology on which they had relied were labeled "behavioral economics" in the press.[2] And so it came to pass that many applications of social and cognitive psychology came to be called behavioral economics, and many psychologists

discovered that the name of their trade had changed even if its content had not. Quite a few of the authors of chapters in this book would be incorrectly described in the press as behavioral economists, because what they do is develop some of the fundamental theories and document some of the central findings on which *Nudge*, and related writings, have relied. This is not the outcome that most researchers, including the authors of *Nudge*, encouraged or viewed as desirable. Richard Thaler has always insisted on a narrow definition of behavioral economics as a distinctive approach of economics, and he would prefer to see "nudges" described as applications of behavioral science.

Labels matter, and the mislabeling of applied behavioral science as behavioral economics has consequences. Some are positive; behavioral economics has retained the cachet of economics, so psychologists who are considered behavioral economists gain some credibility in the policy and business worlds. But the cost is that the important contributions of psychology to public policy are not being recognized as such, and there is the very real worry that young psychologists will be put off from doing policy-related work because they do not consider themselves economists, even with the modifier "behavioral" as a prefix. It is regrettable that the discipline of psychology gets no credit for the most consequential applications of psychological wisdom, and that students of psychology, who ought to take greater pride in their profession, are left to wonder about the contributions of their discipline to society.

In fact, there is a lot to be done. Nudges are an effective way to use psychological insight in the design of policies that might generate greater welfare. But some policy issues will need a greater rethinking: a questioning of the fundamental assumptions, rather than nuanced design. When it comes to the memories of eyewitnesses, or to employers' ability to avoid discrimination, or to the budgeting challenges of the poor, behavioral research presents the serious possibility that we may want to rethink some fundamental concepts and question the basic assumptions of current policies—in other words, do more than merely nudge.

I hope this book helps steer us in the right direction in giving behavioral scientists a greater role in policy making around the world. The chapters of this book, written predominantly by psychologists, illustrate how much psychology has to offer to policy. An important conclusion that readers should draw from it is that modern psychology has agreed on some important aspects of both human nature and the human condition. Recent years have seen a convergence of views on the roles of cognitive and emotional factors as determinants of behavior—and therefore as targets for policy interventions that are proposed to modify people's circumstances or their actions. There is also a growing recognition of the role of social and cultural drivers of behavior, though many social scientists will still complain that psychology is insufficiently attuned to issues of culture and identity. The recognition of the huge power of situation, context, priming, and construal is common ground. We are all Lewinians now, and in the context of policy behavioral economists are Lewinian as well.

The relationship between psychology and economics in the domain of policy was a central issue when psychology became one of the core disciplines in the Woodrow Wilson School at Princeton. In a very different way it is still a dilemma, not because the disciplines are so alien, but rather because they are so close. The overlap of interests and methods is much greater than it was fifteen years ago. Indeed there are several domains in which members of the different tribes deal with similar problems in similar ways. The study of happiness is one of these domains, the study of inequality and poverty may be another. And there will be more. We need a common label for our shared activities. "Behavioral economics" is not a good label, simply because psychologists are not economists and are not trained to think about markets. "Social psychology" would cause similar difficulties to the economists, lawyers, and physicians who engage in Lewinian practice. A descriptively correct label is "applied behavioral science." I would be proud to be called an applied behavioral scientist, and I believe most of the authors of this book would also be happy to be counted as members of this club. This book is a fine illustration of the potential contribution of applied behavioral science to policy.

Notes

1. "We" refers to myself, Eldar Shafir, and Rob McCoun, who came to help us from the Goldman School of Public Policy at Berkeley.

2. Not only in the popular press. I am on record as describing *Nudge* as "the major accomplishment of behavioral economics." I was quite slow to recognize the problem that I address in this foreword.

Contributors

Adam L. Alter

Assistant Professor of Marketing, Leonard N. Stern School of Business, New York University

Adam Alter's research focuses on decision making and social psychology, with a particular interest in the effects of subtle environmental cues on human cognition and behavior.

John P. Balz

Senior Planner, Draftfcb

John Balz develops marketing strategies for major global firms. He was introduced to behavioral ideas in graduate school, where his research focused on behavioral decision making in campaigns and elections, congressional lobbying, and ideological attitude formation.

Mahzarin R. Banaji

Richard Clarke Cabot Professor of Social Ethics, Department of Psychology, Harvard University

Mahzarin Banaji is interested in the unconscious assessment of self and others, its roots in feelings and knowledge about social group membership, and the perceived distinction between "us" and "them" that is often the result.

Michael S. Barr

Professor of Law, School of Law, The University of Michigan

Michael Barr conducts large-scale empirical research on financial services and low- and moderate-income households and writes about a wide range of issues in financial regulation and the law.

Shlomo Benartzi

Professor, Anderson Graduate School of Management, UCLA

Shlomo Benartzi's interests lie in behavioral finance with a special interest in personal finance and participant behavior in defined contribution plans.

Paul Brest

President, William and Flora Hewlett Foundation; Professor Emeritus and Former Dean, Stanford Law School

Paul Brest has most recently focused on issues of problem solving, judgment, and decision making in the context of the law.

Geoffrey L. Cohen

James G. March Professor of Organizational Studies in Education and Business; Professor, Department of Psychology, Stanford University

Geoffrey Cohen's research examines processes related to identity maintenance and their implications for social problems.

William J. Congdon

Research Director, Economic Studies Program, The Brookings Institution

William Congdon studies how best to apply behavioral economics to public policy.

John M. Darley

Warren Professor of Psychology, Department of Psychology, Princeton University

John Darley's research focuses on moral decision making, decisions to punish another for a transgression, and how people attempt to manage others through incentive systems.

John F. Dovidio

Professor, Department of Psychology, Yale University

John Dovidio investigates issues of social power and social relations, with a focus on conscious and unconscious influences on how people think about, feel, and behave toward others based on group membership.

Sara L. Eggers

Independent Researcher

Sara Eggers' work focuses on risk communication, stakeholder engagement, and the development and use of decision-analytic tools to support policy making in the environmental and public health arenas.

Phoebe Ellsworth

Frank Murphy Distinguished University Professor of Law and Psychology, The University of Michigan

Phoebe Ellsworth's research interests lie at the crossroads of psychology and law, particularly person perception, emotion and public opinion, the death penalty, and jury behavior.

Baruch Fischhoff

Howard Heinz University Professor, Department of Social and Decision Sciences, Department of Engineering and Public Policy, Carnegie Mellon University

Baruch Fischhoff's research looks at issues around risk analysis, communication and management, adolescent decision making, informed consent, and environmental protection.

Susan T. Fiske

Eugene Higgins Professor of Psychology, Department of Psychology, Princeton University

Susan Fiske's research addresses how stereotyping, prejudice, and discrimination are encouraged or discouraged by social relationships, such as cooperation, competition, and power.

Craig R. Fox

Professor of Policy, Andersen School of Management; Professor of Psychology, Department of Psychology, UCLA

Craig Fox studies behavioral decision theory, with a focus on how people make judgments and decisions under conditions of risk, uncertainty, and ambiguity.

Julio Garcia

Research Associate, Department of Psychology and Neuroscience, University of Colorado, Boulder

Julio Garcia works with schools to examine processes related to identity maintenance with the aim to reduce racial and gender gaps and improve academic performance.

Alan S. Gerber

Charles C. and Dorathea S. Dilley Professor of Political Science, Yale University

Alan Gerber studies the application of experimental methods to the study of campaign communications. He has designed and performed experimental evaluations of partisan and nonpartisan campaigns and fundraisers.

Daniel G. Goldstein

Principal Research Scientist, Yahoo Research

Daniel Goldstein studies heuristics and bounded rationality as they relate to matters of business and policy, with an emphasis on perceptions of risk and uncertainty, particularly in financial markets.

Ryan Goodman

Anne and Joel Ehrenkranz Professor of Law, New York University School of Law

Ryan Goodman's interests lie in the areas of international human rights law and international relations, focusing on the evaluation of human rights treaties and international law more generally.

Sam Gross

Thomas and Label Long Professor of Law, School of Law, The University of Michigan

Sam Gross' work focuses on the death penalty, false convictions, racial profiling, eyewitness identification, the use of expert witnesses, and the relationship between pretrial bargaining and trial verdicts.

Curtis D. Hardin

Professor, Department of Psychology, CUNY, Brooklyn College

Curtis Hardin's research focuses on the interpersonal foundations of cognition, including the self-concept, social identification, prejudice, and ideology.

Derek Jinks

The Marrs McLean Professor in Law, School of Law, The University of Texas at Austin

Derek Jinks' research and teaching interests include public international law, international humanitarian law, human rights law, and criminal law.

Christine Jolls

Gordon Bradford Tweedy Professor of Law and Organization, Yale Law School

Christine Jolls' research and teaching concentrate in the areas of employment law, privacy law, behavioral law and economics, and government administration.

Leslie John

Assistant Professor of Business Administration, Harvard Business School

Leslie John uses both laboratory and field studies to investigate behavioral questions at the intersection of marketing and public policy concerning consumer privacy and well-being.

Eric J. Johnson

Norman Eig Professor of Business Marketing, Columbia Business School

Eric Johnson's research interests are in consumer and managerial decision making and electronic commerce.

Linda H. Krieger

Professor of Law, William S. Richardson School of Law, University of Hawai'i at Mānoa

Linda Krieger's research focuses on disability discrimination, affirmative action, law and social cognition, judgment in legal decision making, and theories of law and social change.

Howard Kunreuther

James G. Dinan Professor of Decision Sciences and Business and Public Policy, The Wharton School, University of Pennsylvania

Howard Kunreuther's interests involve the ways that society can better manage low-probability/high-consequence events related to technological and natural hazards.

Judith Lichtenberg

Professor of Philosophy, Georgetown University

Judith Lichtenberg focuses on ethics and political philosophy, with special interests in justice and charity, race and ethnicity, and moral psychology.

George Loewenstein

Herbert A. Simon Professor of Economics and Psychology, Department of Social and Decision Sciences, Carnegie Mellon University

George Loewenstein's work brings psychological considerations to bear on models and problems that are central to economics, with a special interest in intertemporal choice.

Elizabeth F. Loftus

Distinguished Professor of Social Ecology, and Professor of Law and Cognitive Science, University of California, Irvine

Elizabeth Loftus studies human memory and examines how facts, suggestions, and other postevent information can modify people's memories.

Robert Meyer

Gayfryd Steinberg Professor of Marketing, The Wharton School, University of Pennsylvania

Robert Meyer's research focuses on consumer decision analysis, sales response modeling, and decision making under uncertainty.

Erwann Michel-Kerjan

Adjunct Associate Professor, Operations and Information Management Department, and Managing Director, Risk Management and Decision Processes Center, The Wharton School, University of Pennsylvania

Erwann Michel-Kerjan's work focuses on strategies for managing the risks, the financial impact, and the public policy challenges associated with catastrophic events.

Dale T. Miller

The Class of 1968 / Ed Zschau Professor of Organizational Behavior, Stanford Graduate School of Business

Dale Miller's research interests include the impact of social norms on behavior and the role that justice considerations play in individual and organizational decisions.

Sendhil Mullainathan

Professor, Department of Economics, Harvard University; Scientific Director and Founder, ideas42

Sendhil Mullainathan conducts research on development economics, behavioral economics, and corporate finance, with a focus on the application of the behavioral perspective to policy.

Ehud Peleg

Head of Enterprise Risk Management, Bank Leumi

Ehud Peleg's interests are in the applications of behavioral science to investment and resource allocation decisions at the individual executive, committee, and board levels.

Deborah A. Prentice

Alexander Stewart 1886 Professor of Psychology, Department of Psychology, Princeton University

Deborah Prentice studies how norms guide and constrain behavior, how people respond when they feel out of step with prevailing norms, and how they react to those who violate social norms.

Emily Pronin

Associate Professor, Department of Psychology, Princeton University

Emily Pronin studies the asymmetries in how we perceive ourselves versus how we perceive others and the ways in which the underlying processes can lead to misunderstanding and conflict.

Donald A. Redelmeier

Canada Research Chair in Medical Decision Sciences; Professor of Medicine, University of Toronto; Director of Clinical Epidemiology, Sunnybrook Health Sciences Centre.

Donald Redelmeier studies decision sciences in medical contexts, particularly the role of judgmental error and the opportunity for improvement in policy and treatment.

Jennifer A. Richeson

Professor, Department of Psychology, Northwestern University

Jennifer Richeson studies the ways in which social group memberships such as race, socioeconomic

status, and gender impact prejudice, stereotyping, and intergroup relations.

Todd Rogers
Assistant Professor of Public Policy, Harvard Kennedy School
Todd Rogers uses the tools and insights of behavioral science to understand and study how to influence socially consequential problems.

Lee Ross
Professor, Department of Psychology, Stanford University
Lee Ross's research focuses on the biases that lead people to misinterpret each other's behavior, thereby creating systematic barriers to dispute resolution and the implementation of peace agreements.

Kathleen Schmidt
Graduate Student, Department of Psychology, The University of Virginia
Kathleen Schmidt's research focuses on social cognition, specifically empathy, racial bias, and how social and environmental feedback can influence self-perception.

Eldar Shafir
William Stewart Tod Professor of Psychology and Public Affairs, Princeton University; Scientific Director and Founder, ideas42
Eldar Shafir's research centers around judgment and decision making and issues related to behavioral economics, with a special interest in poverty and the application of behavioral research to policy.

J. Nicole Shelton
Professor, Department of Psychology, Princeton University
Nicole Shelton's research focuses on understanding prejudice and discrimination from the target's perspective.

Paul Slovic
Professor, Department of Psychology, University of Oregon; President, Decision Research Group
Paul Slovic studies the influence of risk-perception and affect on decisions concerning the management of risk in society, with a special interest in the psychological factors that contribute to apathy toward genocide.

Nancy K. Steblay
Professor, Psychology Department, Augsburg College
Nancy Steblay's research interests are in psychology and law, specifically eyewitness accuracy, pretrial publicity, and inadmissible evidence and judicial instructions to disregard it.

Cass R. Sunstein
Felix Frankfurter Professor of Law, Harvard Law School; Administrator of the White House Office of Information and Regulatory Affairs
Cass Sunstein's research involves the relationship between law and human behavior and particularly focuses on constitutional law, administrative law, environmental law, and law and behavioral economics.

Richard H. Thaler
Ralph and Dorothy Keller Distinguished Service Professor of Behavioral Science and Economics, The University of Chicago Booth School of Business
Richard Thaler studies behavioral economics and finance as well as the psychology of decision making as it influences the conversation between economics and psychology.

Tom Tyler
Professor, Department of Psychology, New York University
Tom Tyler's research is concerned with the dynamics of authority within groups, organizations, and societies and focuses on factors that shape people's motivations when dealing with others in group settings.

Peter Ubel
John O. Blackburn Professor of Marketing, Fuqua School of Business; Professor, Sanford School of Public Policy, Duke University
Peter Ubel's research explores the role of values and preferences in health care decision making, including topics like informed consent, shared decision making, and health care rationing.

Kevin G. Volpp
Professor of Medicine and Health Care Management, The Wharton School, University of Pennsylvania
Kevin Volpp focuses on developing and testing innovative behavioral economics applications to improve patient health behavior and affect provider performance.

Brian Wansink
John Dyson Professor of Marketing and Nutritional Science, Cornell University
Brian Wansink researches food-related consumer behaviors, with a focus on "mindless eating," the study of how microenvironments influence what and how much people eat.

Elke U. Weber

Jerome A. Chazen Professor of International Business; Professor of Psychology and Earth Institute Professor, Columbia University

Elke Weber's research focuses on decision making under uncertainty and on individual differences in risk taking and discounting, specifically in risky financial situations and around environmental decisions.

Andrew K. Woods

Climenko Fellow; Lecturer on Law, Harvard Law School

Andrew Woods's research interests include international human rights and criminal law, with a particular focus on interdisciplinary approaches to law and policy.

David Zionts

Special Advisor to the Legal Adviser, U.S. Department of State

David Zionts's research interests include international law, human rights, and U.S. foreign relations and national security law.

Acknowledgments

This volume is the outcome of a great collaboration, among academic researchers, practitioners, supportive institutions, dedicated staff, funders, and thought leaders who saw the value in this endeavor and hoped that it will help improve policy thinking in years to come. Eric Wanner and the Russell Sage Foundation have been early and consistent supporters of behavioral research into policy and also supported the present project in its early stages. As dean, Michael Rothschild first introduced a behavioral component to the teaching and research at the Woodrow Wilson School of Public and International Affairs, and the subsequent deans, Anne-Marie Slaughter, Nolan McCarty, and Christina Paxson continued to provide constant support thereafter. Princeton's Department of Psychology, its Langfeld Fund, and Deborah Prentice, the department's chair, all provided great support and encouragement.

In addition to several authors who also helped with the reviewing process, others who helped review and improve the contributions to this book include Bob Cialdini, Frank Dobbin, Shane Frederick, Tom Gilovich, Richard Leo, Anastasia Mann, Danny Oppenheimer, Betsy Levy Paluck, Donald Redelmeier, Dan Simon, and Marian Wrobel. Several students and assistants, including Alexandra Cristea, Izzy Gainsburg, Maia Jachimowicz, Lily Jampol, Marion Kowalewski, David Mackenzie, Ani Momjian, Amy Ricci, Jeremy Spiegel, and Abby Sussman, provided great logistical support during various stages of this long project. Seth Ditchik of the Princeton University Press was invaluable in helping conceive of this project and, along with Janie Chan, Beth Clevenger, and Gail Schmitt, helped get it through to its beautifully finished form.

The Behavioral Foundations of Public Policy

Introduction

ELDAR SHAFIR

If you look in the dictionary under *policy*, *public policy*, or *social policy*, you find definitions that amount to the following: a system of regulatory measures, laws, principles, funding priorities, guidelines and interventions promulgated by a person, group, or government for the changing, maintenance or creation of living conditions that are conducive to human welfare. Mostly what these measures, laws, principles, and interventions are intended to do is to shape society in desirable ways: to promote behaviors that yield outcomes conducive to human welfare. Successful policy, therefore, must depend on a thorough understanding of human behavior. What motivates and incentivizes people when they snap into action as opposed to procrastinate, obey or disobey the law, understand or misunderstand, act or fail to act on their intentions, care or do not care, attend or get distracted? How do they perceive their decisions and the options at their disposal? How do they think about what others are doing? These are all questions that must be addressed for the design and implementation of policies to prove successful.

In light of the centrality of behavioral assumptions to policy, it is remarkable how small a role the attempt to understand human behavior has played in policy circles, as well as in the social sciences more generally. It is particularly remarkable because, as we have now come to understand, much of our intuition about human behavior fails to predict what people will do. And policies based on bad intuitive psychology are less likely to succeed and can often prove hurtful. As the economist John Maurice Clark pointed out nearly a century ago, if the policy maker does not seriously study psychology, "he will not thereby avoid psychology. Rather, he will force himself to make his own, and it will be bad psychology" (*Journal of Political Economy*, 1918).

Bad psychology comes in many forms. A naive understanding of incentives, for example, might suggest that paying people some small amount (rather than nothing) to perform a societally desirable act could only increase instances of that act; instead, it turns out that the loss of the "psychic" benefit of having been a good citizen (which is largely neutralized by the monetary remuneration) can, in fact, reduce take-up. Alternatively, presenting lineups (where suspects are observed concurrently) versus show-ups (where they are seen one at a time) may appear normatively indistinguishable, but we now know that the former leads to more false identifications than the latter. Similarly, having workers opt out of, rather than opt into, retirement savings accounts, looks like an immaterial nuance, except that the former, for predictable reasons and for what amounts to very little cost, generates many more happy retirees than the latter.

A careful consideration of the role of psychology in public policy took many years to develop even after Clark's warning about the dangers of bad psychological assumptions. An important turning point was the behavioral critique of the economic assumptions underlying individual decision making begun by cognitive and social psychologists in the 1970s. This was eventually reinforced by the economic profession's gradual, even if reluctant, acceptance of the behavioral critique and led to increased research applying behavioral insights to studies of choice and judgment in everyday life. Now, almost a half century after the emergence of the modern critique, the behavioral perspective occupies a respectable and increasingly popular niche in many graduate programs in economics, business, law, policy, and the social sciences more generally. And thus we have arrived at a point where it is only natural to explore how best to incorporate elements of the behavioral perspective into policy thinking.

The behavioral findings provide an alternative view of the human agent. Many aspects of decision making that the normative analysis assumes do not matter (such as how the options are described, as long as the same information is given) prove highly consequential behaviorally, and other factors that are normatively assumed to be of great importance (such as whether an intervention will help save 1,000 birds or 10,000 birds) are, instead, intuitively largely ignored. At the most general level, a couple of deep lessons have emerged that are of great potential relevance to policy makers: the relevance of context and the unavoidability of construal.

Human behavior tends to be heavily context dependent. One of the major lessons of modern psychological research is the impressive power that the situation exerts, along with a persistent tendency on our part to underestimate this power relative to the presumed influence of personal intentions and traits. In his classic obedience studies, for example, Milgram (1974) demonstrated the ability of decidedly mild situational factors to trigger behaviors on the part of regular citizens, such as the administration of presumed electric shocks to innocent others, that were unfathomable to naive introspection. Along similar lines, Darley and Batson (1973) showed how seminary students on their way to deliver a sermon on the parable of the Good Samaritan were willing to walk right past an ostensibly injured man, slumped coughing and groaning, simply because they were running late. Minor contextual features were shown in these cases to override people's professed intentions and their deeply held convictions. To the extent that such minor contextual features are able to transcend education, personality, and intention, policy makers appear to have powers of influence that they underappreciate, may unintentionally misuse, and could, given some behavioral insight, employ better.

The second lesson, which is fundamental to the cognitive sciences in general, concerns the role of "construal" in mental life. People do not respond to objective experience; rather, stimuli are mentally construed, interpreted, and understood (or misunderstood). While this claim risks sounding deep, it is actually trivial, but with profound consequences: behavior is directed not toward actual states of the world but toward mental representations of those states, and those representations do not bear a one-to-one correspondence with the states they represent. In fact, the representations we construct may not even constitute faithful renditions of actual circumstances. Our visual experience, for example, is the product of complex processes that take raw visual input (say, a trapezoid when we look at a window from the side) and use contextual cues to represent what is "really there" (a perfectly rectangular window). Anytime those cues are misleading, we end up with a false representation, as in the case of well-known optical illusions. How we interpret attitudes and emotions is similarly a matter of construal. And, as it turns out, so is our representation of many objects of judgment and preference. We can only decide between things as they are represented in the three-pound machine that we carry behind the eyes and between the ears. And those representations are the outcome of mental processes that, to some extent at least, have a life of their own.

For policy makers all this should be of the utmost importance. Policies' success depends on human behavior. And behavior is determined not simply by what is available, but by what people know, perceive, understand, attend to, and want. Thus, well-intentioned interventions can fail because of the way they are construed by the targeted group. And the difference between success and failure can sometimes boil down to a relatively benign and normatively immaterial change in presentation and construal, rather than a complex and costly rearrangement of the available alternatives.

About fifteen years ago, we began a joint formal program of training in "psychology for policy" between the psychology department and the Woodrow Wilson School of Public and International Affairs at Princeton University. The endeavor was new at the time, and the results of the initiative were not entirely predictable. What were some of the more important policy questions to which a behavioral analysis could significantly contribute? Where in policy did misguided behavioral assumptions figure most prominently, where were they of lesser importance, and what exactly were they anyway? How was one to go about researching and communicating all this? And would it make a difference? As often happens when ideas gather momentum, we were not alone. An increasingly talented and interdisciplinary group of scholars had grown interested in research along similar lines and in issues of both behavioral and policy significance.

The present volume presents some of the more impressive outcomes of this important work, as conceived and summarized by many of the leading scholars in the field. The wide array of topics covered here should appeal to students of human behavior interested in real-world applications. More importantly, the chapters in this volume were prepared with an eye toward a sophisticated audience with no behavioral training. The application of experimental findings and concepts emanating from behavioral research to the design and implementation of policy—call it "behavioral policy"—is an exciting and rapidly expanding new area of research and study. The present collection is intended to expose policy makers, practitioners, government officials, business leaders, legal, ethical, and health professionals, as well as students interested in societal, domestic, and international challenges, to a perspective that can shed new light. Greater insight into human behavior, the authors in this volume agree, can prove helpful, both in making sense of what are otherwise persistent puzzles, as well as in generating novel ideas and effective solutions.

The contributions to this collection tend to be highly interdisciplinary and thus hard to compartmentalize. Nonetheless, the sheer amount of material presented in this volume warrants some minimal organization in the hopes of facilitating the reader's task. Chapters are divided by general topics but are

otherwise independent and can be read in any order; occasional cross-references occur when the materials of separate chapters are especially complementary. The aim of each chapter is to provide the reader with an overview of how research in the behavioral sciences might influence our understanding and the conduct of good policy in a particular domain. Ultimately, we hope the reader will come to see the foundational role that behavioral assumptions must come to play in shaping the successful design and implementation of policy.

The Chapters

The early chapters focus on behavioral issues that arise in the conduct of our social and political lives. They focus on policy-relevant topics ranging from the nature of intuitive social judgment and "automatic" social perceptions to the valuation of social belonging and concerns with identity, justice, and fairness; problems ranging from discrimination in the work place to the numbing that comes with hearing about mass atrocities.

A common thread running through these contributions is that the empirical findings are often in tension both with normative assumptions as well as with common intuition. As a result, they have far-reaching consequences for how we think about policy design and implementation. We tend to think, for example, that people's behavior largely mirrors their beliefs and that their choices are typically about tangible, value-maximizing outcomes. Thus, we might assume, those who are not prejudiced will typically not exhibit prejudiced judgment, and if voting is unlikely to have a tangible impact, people will not bother to vote. Similarly, the intuition goes, negotiators whose persistent biases lead to impasse will learn to overcome them, and managers whose unintended discriminatory practices hurt their business will learn to avoid discriminating.

In contrast to all that, as the chapters below illustrate, people care a lot about intangibles, they exhibit persistent biases in social perception, and they lack introspective access to the biases and the motivations that often are in tension with their better judgment. As a result, people often fail to recognize the discrepancies between their beliefs and their actions, which, rather than resolving themselves in the long run, often end up playing a big role in exacerbating longstanding political and social tensions.

Prejudice and Discrimination

In the opening chapter, on implicit prejudice, **Curtis Hardin** and **Mahzarin Banaji** argue that our views of prejudice and discrimination are based on outdated notions, with important policy implications. Rather than arising from ignorance and hatred, which would be best addressed by changing the hearts and minds of individuals, prejudice and stereotyping, according to these authors, emerge from cognitively salient structures in our social milieu and do not necessarily involve personal animus, hostility, or even awareness. Rather, prejudice is often "implicit"—that is, unwitting, unintentional, and uncontrollable—even among the most well intentioned. At the same time, these authors suggest, research shows that implicit prejudice can be reduced through sensible changes in the social environment.

The social environment figures prominently in **Nicole Shelton**, **Jennifer Richeson**, and **John Dovidio**'s chapter on intergroup biases in interracial interactions. The goal of this chapter is to explore how racial bias can influence affective, cognitive, and behavioral outcomes during interracial interactions, especially among those who do not harbor explicitly racist attitudes. This question is examined in a variety of contexts, including students sharing dorm rooms on university campuses and interactions between White physicians and racial minority patients in health care settings. A central message that emerges from these interactions is that bias is expressed in subtle ways: as strained relationships between roommates, as less effective interactions between physicians and patients, and as lower levels of rapport in employment interviews. In each of these cases, there is rarely an obvious act of blatant discrimination. Instead, the complex and often subtle nature of contemporary intergroup bias, for which traditional policies designed to respond to overt discrimination are ill suited, can have widespread impacts on intergroup interactions, often with different consequences for members of different racial and ethnic groups. Shelton et al. conclude with a review of common practices and interventions that policy makers could use to maximize the benefits of diversity across policy-relevant settings.

In their chapter on gender bias, **Susan Fiske** and **Linda Krieger** consider the legal ramifications of unexamined gender discrimination, particularly as it plays out in employment contexts. They review recent behavioral and neuroscience research that challenges the rational-actor assumption underlying much of the debate over discrimination law and policy. Decision makers, according to Fiske and Krieger, cannot always make optimal employment decisions, because, even when they consciously support equal opportunity norms, subtle, unexamined forms of gender bias may prevent them from accurately perceiving decision-relevant information, or from optimally using it to make employment decisions. Managers may explicitly endorse equal opportunity, but unexamined prejudices

might nevertheless derail their choices. Fiske and Krieger consider the kinds of initiatives that organizations might undertake in an attempt to reduce the levels of workplace discrimination caused by unexamined, subtle bias. They also advocate for policy initiatives, including a mandatory information disclosure approach to equal opportunity employment policy, which they suggest might help squeeze discrimination out of labor markets in ways that circumvent the need to identify individual instances of discriminatory decision making.

Social Interactions

In his chapter on the psychology of cooperation, **Tom Tyler** argues that, while incentives and sanctions matter, standard normative approaches place too much emphasis on issues of material gains and losses. Tyler analyzes laboratory and field studies that illustrate several types of social motivations—attitudes, values, personal identity, procedural justice, and motive-based trust—that have a strong influence on behaviors in social settings. Tyler focuses on the problem of motivating cooperative behavior and suggests that policy makers have a great deal to gain from expanding their conception of human nature and recognizing the importance of social motivations in shaping people's behavior in groups and organizations.

Along related lines, **Todd Rogers**, **Craig Fox**, and **Alan Gerber** propose an alternative conceptualization for why people vote. Rather than the standard self-interested view, which cannot explain the decision to vote given the miniscule probability that one's vote will affect the outcome, the authors propose to think of voting as a "dynamic social expression." Voting, according to this perspective, is the outcome of a dynamic constellation of events that extend over time; it is an inherently social act, and it is ultimately an expression of one's identity. Among other things, Rogers, Fox, and Gerber describe recent experimental field research into effective get-out-the-vote campaigns, thereby linking the question of why people vote to an array of behavioral research—including social and cognitive psychology and behavioral economics—that has not been systematically linked to voting behavior in the past.

In his chapter on disagreement, **Lee Ross** considers several cognitive and motivational processes and the role they play in adding hostility and distrust to policy disagreements, and how they serve as barriers to dispute resolution. Among other constructs, he considers the *reactive devaluation* of proposals put forth by the other side and the role of *naive realism*, the conviction that one sees things objectively and clearly, which tends to add rancor to disagreement insofar as it creates the expectation that other reasonable and honest people ought to share one's views. (This perspective was well captured by comedian George Carlin's observation about driving: "Ever notice that anyone going slower than you is an idiot and anyone going faster is a maniac?") Informed by the foregoing analysis, Ross then considers several behaviorally informed strategies for overcoming barriers to agreement.

Finally, in their chapter on psychic numbing, **Paul Slovic**, **David Zionts**, **Andrew Woods**, **Ryan Goodman**, and **Derek Jinks** ask why people repeatedly fail to react to genocide and other mass-scale human atrocities. It is not, they argue, because people are insensitive to the suffering of their fellow human beings, or even that most only care about identifiable victims of similar skin color who live nearby. Rather, they suggest, a fundamental problem lies in people's incapacity to experience commensurate *affect*, the positive and negative feelings that combine with reasoned analysis to guide action. Left to its own devices, moral intuition appears to respond more to individual stories that are easier to imagine than to statistics of mass atrocities, which fail to spark commensurate affect and motivate appropriate action. Even when we know genocide is real, we do not "feel" that reality. The authors explore some behaviorally informed ways that might make genocide "feel real," but they are ultimately led to the conclusion that we cannot rely on intuitive reactions and must instead commit ourselves to institutional, legal, and political responses that are less susceptible to psychic numbing and more heavily based upon reasoned analysis of our moral obligations.

The Justice System

The rational agent model has figured prominently in many areas of the law. At the same time, much of what comes under the law depends on the impulses, intuitions, judgments, sense of confidence, emotional reactions, and everyday understandings of regular citizens when they act as witnesses, jurors, colleagues, employers, employees, and so forth. And because the legal system is heavily in the business of constructing rules and procedures, there is much room to think about how these can be better shaped by a nuanced understanding of human capabilities, proclivities, and limitations.

In their chapter on eyewitness identification and the legal system, **Nancy Steblay** and **Elizabeth Loftus** focus on issues of eyewitness memory, such as the fact that faulty eyewitness memory has been implicated in a majority of (mostly DNA-based) exonerations. They review the main lessons from the science of eyewit-

ness memory and consider their implications for improving the legal system whenever eyewitnesses are playing a crucial role. They provide a primer on the essential memory principles underlying eyewitness performance, including the fact that this experience is not just a memory phenomenon, but that it also reflects social forces, including, for example, subtle and unintentional verbal and nonverbal communications from others. What emerges is the potential for memory to be contaminated, distorted, and yet reported with great confidence, which proves of great relevance to a legal system that depends on and believes in eyewitness veracity and in which many people become criminal defendants on the basis of eyewitness identification. Steblay and Loftus describe the ongoing research effort around the topic of eyewitness testimony, the changes in legal policy spurred by the collaboration between behavioral scientists and those in the legal field, and the challenges that persist in the application of memory research to public policy.

In "False Convictions," **Phoebe Ellsworth** and **Sam Gross** extend the analysis of the rate and persistence of false convictions to a variety of psychological, social, and institutional factors beyond eyewitness identification. They first highlight the inherent difficulty of detecting false convictions, where the only ones we know of (and even there we could be wrong) are exonerations: those rare cases in which a convicted criminal defendant—typically in the most serious of cases, the only ones to receive sufficient attention and resources—is able to prove his innocence after the fact. Ellsworth and Gross consider the social and institutional context that characterizes criminal investigation and adjudication under the adversarial system. They analyze the proclivity of the process to give rise to worrisome behavioral phenomena, including confirmation biases, eyewitness misidentification, false confessions, fraud and error on the part of forensic analysts, perjury by jailhouse informants and other witnesses, misconduct by police and prosecutors, and incompetent representation by criminal defense attorneys. Ellsworth and Gross describe the relevant work by social scientists and legal researchers and consider some areas for future policy enhancement.

In a chapter focusing on behavioral reactions to wrongdoing, **John Darley** and **Adam Alter** explore the nature and consequences of potential gaps between legal codes and community sentiments regarding punishment, retribution, and deterrence. They first review research on people's perceptions of wrongful actions and the punishments those actions deserve. They conclude that people are driven by emotionally tinged reactions of moral outrage and that their punishment decisions are largely based on what they intuitively believe the offender justly deserves. They then consider conventional approaches to dealing with crime, punishment, and deterrence and conclude that in light of what we know about human cognition and behavior, those approaches are largely ineffective. For example, whereas our penal system focuses heavily on sentence duration, sentence duration is generally an ineffective deterrent, as compared, for example, to salient surveillance mechanisms. Darley and Alter consider relevant policy implications, while keeping in mind that citizens' intuitive perceptions of justice will place limits on the types of societal punishment practices that will be perceived as fair and that legal codes that clash with those moral sensibilities can cause citizens to lose respect for the law.

Bias and Competence

Of great relevance to policy are the circumstances in which people exhibit systematic bias or fail to weigh appropriately the factors that matter most to a decision. In other circumstances, people may perform the requisite tasks exceedingly well. This contrast is heightened by the fact that it is often hard for people to anticipate when they might expect bias as opposed to remarkable judgmental acuity. Things that ought not matter from a normative perspective often do, and things that ought to matter often fail to have an impact. The chapters in this section are motivated by the assumption that greater awareness and the proper anticipation of bias and other behavioral limitations may help devise more effective policies.

In their chapter on claims and denials of bias, **Emily Pronin** and **Kathleen Schmidt** explore the far-reaching policy implications of people's perception that their own judgments are relatively free of bias whereas others' judgments are susceptible to it. People's tendency to be blind to their own biases while exaggerating those of others can lead to a range of problems, among which are social conflict, breakdown of negotiations, corruption, and discrimination. Pronin and Schmidt examine the central behavioral underpinnings of this "bias blind spot" and consider potential solutions, including increased awareness, education, and psychologically savvy disclosure requirements, with an emphasis on how to make those solutions psychologically informed and thus more effective.

In "Questions of Competence: The Duty to Inform and the Limits to Choice," **Baruch Fischhoff** and **Sara Eggers** discuss the nature of assumptions about people's competence that figure, often implicitly, in a wide range of regulatory and policy domains. For example, product disclosure requirements reflect beliefs about people's ability to recruit and comprehend the relevant information, and policies governing

living wills reflect assumptions about our ability to anticipate the relevant circumstances. When competence is underestimated, they argue, people's freedom to make their own decisions may be needlessly curtailed, whereas when competence is overestimated, they may be denied important protections. The chapter considers several applications to risk-related decisions in U.S. policy contexts (including drugs, pathogens, and contaminants) in its attempt to offer a general approach to assessing and, where possible, improving individuals' competence to make the requisite decisions.

In his chapter on misfearing and cost-benefit analysis, **Cass Sunstein** argues in support of cost-benefit analysis as a way of counteracting the problem of misfearing, that is, people's tendency to misperceive risks. Whereas cost-benefit analysis is often justified on conventional economic grounds, as a way of preventing inefficiency, Sunstein argues for it on grounds associated with cognitive and social psychology, including concepts such as availability and salience, informational and reputational cascades, intense emotional reactions, motivated reasoning, and causal misattribution, all of which can lead people to be afraid of fairly trivial risks and neglectful of serious ones. Such misfearing, Sunstein suggests, plays a significant role in public policy because of the power of self-interested private groups and ordinary political dynamics. And when misallocations of public resources result from misfearing and associated problems, cost-benefit analysis can operate as a corrective, a way of ensuring better priority setting and of overcoming behavioral obstacles to desirable regulation.

Behavioral Economics and Finance

Behavioral research in economics and finance has explored the systematic ways in which people's preferences are in tension with standard assumptions underlying the classical theories of choice. Among other things, people tend to focus on perceived departures from the status quo rather than on final assets, they exhibit unstable discount rates, and they tend to be loss averse—the dread of losses is greater than the savoring of equivalent gains. Intangibles such as fairness and inertia matter a lot, and decisions are often made "locally," with much reliance on features that loom large at the moment, often at the expense of long-term objective outcomes. All this puts a greater burden on policy design, since minor and normatively inconsequential changes can make the difference between policies that succeed and those that fail.

This type of analysis, in the context of retirement saving plans, is illustrated by **Shlomo Benartzi**, **Ehud Peleg**, and **Richard Thaler**, who apply behavioral principles to the study of the choices made by employees saving for retirement. Exploring notions ranging from decision inertia and nominal loss aversion to discounting and the synchronization of saving increases with pay raises, they show how supposedly minor details in the architecture of retirement plans can have dramatic effects on investment decisions and savings rates. More generally, they suggest, such insights into the architecture of decision have the potential to help people make better decisions, a theme Thaler returns to in a chapter with Balz and Sunstein later in the book.

Applying a behavioral economic analysis to employment law, **Christine Jolls** considers the lessons of behavioral analysis for legal requirements and rules that govern employer-employee relationships, ranging from wage payment and pension regulation to minimum wages, mandated health insurance, workplace leave, and discrimination laws. The effects of employment law, Jolls argues, turn in significant part on people's behavior in employment settings, which can be illuminated by consideration of bounded willpower, bounded self-interest, and bounded rationality. Thus, errors in intuitive judgment have implications for employment discrimination law, and different rules may prove more effective in encouraging retirement saving by individuals with bounded willpower. Furthermore, a "fairness dynamic"—one in which employers choose to pay employees more than the minimum they would accept and employees respond by working harder than they otherwise would—has implications for minimum wage regulation. The employment relationship, Jolls concludes, is one of life's most important relationships, and it can greatly benefit from a behavioral economic perspective.

In their chapter on decision making and policy in contexts of poverty, **Sendhil Mullainathan** and **Eldar Shafir** present a behaviorally motivated framework for understanding the decisions of the poor. Motivated by empirical insights on judgment and decision making that are supplemented by lessons from social and cognitive psychology, they ask how we might explain behaviors in poverty and how might similar behaviors have different consequences when people are poor. They conclude with recent work in which poverty is viewed as a context that creates unique challenges for the human psyche, above and beyond budgetary woes. Poverty itself, according to Mullainathan and Shafir, generates specific psychological responses that are endemic to functioning with little slack and with constant vigilance and that can lead to distraction, miscalculation, and depletion. This, they propose, suggests new approaches to policy making that are focused on programs that foster stability and give people the financial and psychic steadiness needed to build more robust economic lives.

Behavior Change

Subtle changes in the context of decision can have a significant, and normatively unexpected, impact on the course of action taken. And this has important implications for the familiar tension between intention and action. In the face of contextual obstacles (which can range from transportation to shame to forgetfulness), people can fail to act even when they have a strong intention to do otherwise. In contrast, when the context is designed to facilitate certain actions, those actions might be taken even when resolve is not terribly high. Contextual cues and interventions, incentives, and decision aids, sometimes quite subtle and limited in scope, can have substantial impact even in contexts where preferences otherwise appear clear and strong. This, of course, only increases the onus on policy makers to construct and implement policies that are behaviorally insightful and thus more likely to have the desired effects. It also suggests that at times the passage from a policy that is not working to one that does may require different, and perhaps more nuanced (and affordable), changes, ones that address how people construe a problem and what that construal leads them to do.

In their chapter on the psychological levers of behavior change, **Dale Miller** and **Deborah Prentice** focus on circumstances in which policy makers wish to help people change their behavior in ways that align with these people's own long-term interests and stated wishes. Miller and Prentice analyze the capacity of various interventions to move people toward desirable behavior and away from undesirable behavior, with a special emphasis on the psychological constructs and processes that produce behavior change. Among other things, they illustrate ways in which economic and psychological incentives can combine in complex ways, producing counterintuitive effects from economic taxes and subsidies. They outline how efforts to change behavior must begin with a careful analysis of the motivational dynamics bearing on the status quo and the levers that can be used to change them.

In his chapter, "Turning Mindless Eating into Healthy Eating," **Brian Wansink** considers some of the basic processes behind a variety of environmental factors that influence food consumption. Package size, plate shape, lighting, socializing, and the visibility, variety, size, and accessibility of food are only some of the environmental factors that influence the volume of food consumed and are considered likely to have contributed to an ever-widening obesity problem in many places. Understanding these drivers of consumption volume has immediate implications for nutrition education and consumer welfare, but education and increasing awareness are unlikely to be the solutions, Wansink argues, because the effects occur at perceptual levels of which we are not aware. Instead, he lists some behaviorally informed principles that academics, industry, and government can use when partnering to make tangible health-related changes in the lives of individuals.

In "A Social Psychological Approach to Educational Intervention," **Julio Garcia** and **Geoffrey Cohen** focus on the psychological causes of academic underperformance, particularly the racial achievement gap observed in American schools. Among others, they describe psychological interventions that focus on the presence of an "identity threat" and that when systematically applied have been found to close the achievement gap. At the heart of their analysis is the notion of the classroom as a tension system in which various factors, both structural and psychological, interact to produce an environment that elicits a set of attitudes, behaviors, and performance. By heightening the impact of factors facilitating performance or lessening the impact of factors that impede it, interventions can alter students' psychological environments. This analysis leads Garcia and Cohen to conclude that well-timed interventions targeting important psychological processes can produce effects on performance that appear disproportionately large. Throughout, they discuss the implications for social policy that follow from their approach.

Improving Decisions

Systematic tendencies, ranging from an inadequate weighing of likelihoods to excessive discounting of the future to an inability to simulate future feelings, can all interfere with the making of optimal decisions. Furthermore, limited mental resources and attention have important implications for people's abilities to budget, save, invest in mitigation against natural disasters, or bother to develop a long-term collective perspective. What repeatedly emerges as important is not sheer human ability, which can be impressive, but the fact that intuition, attention, and understanding can be tapped into in ways that are less or more likely to succeed. The contributions that follow consider behaviorally informed ways in which policy makers might help people reach better decisions, individually and collectively.

Going beyond issues of mere comprehension, **Peter Ubel** considers the use of medical decision aids to improve people's "preference sensitive decisions," where the decision maker, when left to her own devices, might not make the right choice. In particular, what is envisioned is a neutral party to help the patient make a decision consistent with her underlying goals and preferences. Experts on decision aids, Ubel argues,

have typically assumed that if you give decision makers full information and the freedom to choose, they will experience reduced conflict and higher satisfaction and will make decisions that reflect their true preferences. Instead, he suggests, decision counselors need to go beyond increased comprehension and conflict reduction, in light of much evidence showing that people who comprehend their options nevertheless can make bad decisions, and that good decisions can still leave people deeply conflicted. Toward that end, Ubel evaluates the strengths and weaknesses of several criteria by which to determine whether a structured decision aid has helped people make good preference-sensitive decisions.

In their chapter, "Using Decision Errors to Help People Help Themselves," **George Loewenstein**, **Leslie John**, and **Kevin Volpp** argue that having identified a variety of systematic decision errors, behavioral researchers are in a good position to provide policy solutions that make use of those same errors to people's benefit. They show how a wide range of decision phenomena that are typically viewed as errors—including the status quo and default bias, loss aversion, overoptimism, and nonlinear probability weighting, among others—can be exploited to help people accomplish their goals, ranging from saving money and losing weight to drug adherence and charitable giving. They also consider whether such errors could be exploited to deal with broader societal problems such as global warming. There are, according to Loewenstein, John, and Volpp, many economic entities that exploit consumers' mistakes. Instead of leaving consumers to fend for themselves, we ought to harness the same errors that are regularly used to exploit them to instead help make people better off.

In her chapter exploring the insights of behavioral decision research for better environmental decisions, **Elke Weber** starts by outlining the logic of environmental policy decisions, which typically include social and economic dimensions, considerations of fairness or equity and considerable uncertainty involving intertemporal tradeoffs and which require foresight, patience, and persuasion. Because environmental goods like clean air, drinkable water, and species diversity are common-pool resources, rational analysis essentially prescribes shortsighted behaviors even if more long-sighted and cooperative solutions are socially desirable. Informed by social cognition and behavioral decision research, Weber argues that insights into unconscious and social inferential and decision processes, as well as into people's limitations in attention, memory, and information processing, can help guide the design of more promising environmental policies. Behaviorally informed considerations, she argues, suggest that people might be induced to act in more collective ways that increase their own long-term benefits, as long as we are able to shape their decision environment in ways that facilitate environmentally sustainable behaviors.

In their chapter on overcoming decision biases to reduce losses from natural catastrophes, **Howard Kunreuther**, **Robert Meyer**, and **Erwann Michel-Kerjan** describe the recent trend of escalating losses from natural hazards. They attribute this to an interplay of economic and behavioral factors: increased levels of assets placed in harm's way often without adequate investments in mitigation, along with a tendency to underattend to low-probability, high-consequence risks and to underappreciate the benefits of long-term investments in protection. The result is an accelerating spiral of risk taking, where the rate of economic development in high-risk areas outpaces investment in technologies intended to protect those developments. Kunreuther, Meyer, and Michel-Kerjan consider some of the behavioral drivers of this mismatch, and how taking these into account might help devise instruments (such as long-term insurance policies coupled with home improvement loans to induce investment in cost-effective mitigation measures) that can help reduce losses from future natural disasters.

Decision Contexts

The concluding three chapters explore several important features of contextual design—defaults, choice architecture more generally, and behaviorally informed regulation—all of which, it is argued, can aid in the implementation of improved policies. In "Decisions by Default," **Eric Johnson** and **Daniel Goldstein** draw on a variety of policy domains to illustrate the power of defaults and then explore some of the psychological mechanisms that underlie these effects. From insurance and organ-donation decisions to retirement savings and internet privacy settings, changing a no-action default can be highly effective compared to economic incentives or extensive educational or persuasion campaigns designed to influence people to make active decisions. Guided by the realization that each kind of default has costs and benefits and by considerations of ethics and effectiveness, Johnson and Goldstein discuss the importance to policy makers of understanding defaults and suggest conditions when different kinds of default arrangements—forced choice; mass defaults; random, smart, or personalized defaults—might be advisable.

In their chapter on choice architecture, **Richard Thaler**, **Cass Sunstein**, and **John Balz** consider decision makers, who—like all of us, if you believe the behavioral findings—function in an environment where many features, noticed and unnoticed, can influence

the decisions that they make. Those who shape the decision environment, in this case the policy makers, are the "choice architects." Thaler, Sunstein, and Balz analyze some of the tools that are available to choice architects, such as creating defaults, expecting errors, understanding mappings, giving feedback, structuring complex choices, and creating incentives. Their goal is to show how choice architecture can be used to help nudge people to make better choices (often as judged by themselves) without forcing the intended outcomes upon anyone, a philosophy they call libertarian paternalism.

Finally, looking at behaviorally informed regulation, **Michael Barr**, **Sendhil Mullainathan**, and **Eldar Shafir** propose a regulatory framework based on insights from behavioral economics and industrial organization in which outcomes are an equilibrium interaction between individuals with specific psychologies and firms that respond to those psychologies within specific market contexts (in contrast to the classic model, which assumes an interaction between rational choice and market competition). The introduction of a richer psychology, Barr, Mullainathan, and Shafir propose, complicates the impact of competition. It suggests that firms compete based on how consumers respond, and competitive outcomes may not always align with increased consumer welfare. Regulation must then address failures in this equilibrium. For example, in some contexts market participants will seek to overcome common human failings (as for example, with undersaving), whereas in other contexts market participants will seek to exploit them (as with overborrowing). Barr et al. discuss specific applications and illustrate, among other things, how a behaviorally informed regulatory analysis could improve policy makers' understanding of the costs and benefits of specific policies.

Commentaries

The volume concludes with a series of commentaries from scholars in four disciplines—philosophy, economics, medicine, and law. These scholars' main lines of research lie outside the behavioral arena, but they all have had a longstanding interest in behavioral applications and took it upon themselves to comment on issues raised in this volume, particularly as they interact with their own disciplinary ways of thinking. **William Congdon** considers some of the ways in which the behavioral perspective can inform economic policy; **Donald Redelmeier** looks at the ways in which behavioral insights might inform health care policy; **Paul Brest** focuses his attention on issues surrounding the potential debiasing of policy makers and lawmakers; and **Judith Lichtenberg** aims a philosophical lens at issues of paternalism, manipulation, and the extent to which behaviorally informed policy making may be good for people.

In the chapters that follow, more than fifty scholars will tell you about a rich body of research conducted over the past three to four decades that has changed the way we understand people. They will consider several implications of the research findings, and they will suggest many ways in which our new understanding, this new view of the human agent, might help design and implement better public policy. We hope that you find this exposition of the behavioral foundations of policy productive and illuminating and that you will use it to create new policies that further improve human welfare.

CHAPTER 1

The Nature of Implicit Prejudice

Implications for Personal and Public Policy

CURTIS D. HARDIN

MAHZARIN R. BANAJI

Some fifty years ago in Arkansas, nine black students initiated a social experiment with help from family, friends, and armed National Guards. Their successful attempt to desegregate Little Rock's Central High School following the decision in *Brown v. Board of Education* is among the most momentous events in America's history, leaving no doubt about its historic importance and the significance of its impact on public policy. Nevertheless, as many have noted, even at the beginning of the twenty-first century, a blatant de facto segregation in living and learning persists and in some circumstances has intensified (e.g., Orfield, 2001). The American experiment in desegregation is a reminder that public policies, however noble in intent, may not realize their aspirations if they do not include an understanding of human nature and culture. In other words, they cannot succeed if they are not founded on relevant scientific evidence, which reveals the nature of the problem, the likely outcomes, and how social transformation can best be imagined. As an example of the importance of basing policy in science, there is the research of Robert Putnam showing the unsavory result that ethnic diversity may actually increase social distrust. As the ethnic diversity by zip code increases, so does mistrust of one's neighbors, even same-ethnicity neighbors (Putnam, 2007). The naive optimism that diversity will succeed in the absence of a clear understanding of the dynamics of social dominance and intergroup relations is challenged by these and other similar revelations (e.g., Shelton, Richeson, and Dovidio, this volume). Hence, even well-intentioned public policies are unlikely to yield positive outcomes unless they are grounded in the best thinking available about how people actually think and behave. Sadly, this has not been the case, both because policy makers are not sufficiently respectful of the importance of science as the guide to social issues and because academic scientists resist imagining the policy implications of their evidence.

In this chapter, we address the topics of stereotyping and prejudice, staying firmly within the bounds of what science has demonstrated. However, in keeping with the mission of this book, we spell out what we see to be some obvious, and also some less obvious, tentacles to questions of public policy. We posed the following questions to ourselves: What are the broad lessons learned that have changed our understanding of human nature and social relations in recent decades? In what way does the new view run counter to long-held assumptions? How should policy involving intergroup relations proceed in light of these discoveries? And, can we speak about "personal policies" that may emerge from the education of individuals about the constraints and flexibility of their own minds while also considering the notion of policy in the usual "public" sense? Our contention is that personal and public policy discussions regarding prejudice and discrimination are too often based on an outdated notion of the nature of prejudice. Most continue to view prejudice as it was formulated generations ago: negative attitudes about social groups and their members rooted in ignorance and perpetuated by individuals motivated by animus and hatred. The primary implication of the old view was that prejudice is best addressed by changing the hearts and minds of individuals, for good-hearted people will think well of others and behave accordingly. However, research in recent years demonstrates that the old view of prejudice is incomplete, even dangerously so. Staying with it would lead to policy choices that might be ineffectual, or worse. Staying with it would be akin to ignoring the evidence on smoking and cancer.

How has the scientific understanding of prejudice changed? In short, we now know that the operation of prejudice and stereotyping in social judgment and behavior does not require personal animus, hostility, or even awareness. In fact, prejudice is often "implicit"—that is, unwitting, unintentional, and

uncontrollable—even among the most well-intentioned people (for a review, see Dovidio and Gaertner, 2004). Moreover, although the discovery of implicit prejudice initially brought with it an assumption that it might be unavoidable (e.g., Bargh, 1999; Devine, 1989; Dovidio et al., 1997), research demonstrates that, although it remains stubbornly immune to individual efforts to wish it away, it can be reduced and even reversed within specific social situations through sensible changes in the social environment (e.g., Lowery, Hardin, and Sinclair, 2001; Rudman, Ashmore, and Gary, 2001). In sum, in addition to the real problems that malicious "bad apples" pose for social policy, research demonstrates that prejudice also lives and thrives in the banal workings of normal, everyday human thought and activity. In fact, an overemphasis on the bad apples may well be detrimental to considerations of policy because it assumes the problem of prejudice to be that of the few rather than that of the many (Banaji, Bazerman, and Chugh, 2003).

We believe that the new understanding of prejudice that has evolved over the past three decades invites a transformation of the public debate regarding how the problem of prejudice may be productively addressed. Hence, this chapter will review the research that has so dramatically changed the contemporary understanding of the nature of prejudice, with an emphasis on research demonstrating (a) the existence of implicit prejudice, (b) the ubiquity of implicit prejudice and its consequences, (c) principles by which the operation of implicit prejudice may be influenced, and (d) the policy changes implied by a recognition of what the mind contains and is capable of. In so doing, we argue that although implicit prejudice has disturbing consequences for social judgment and behavior, potential solutions may arise in part from a reconceptualization of prejudice—less as a property of malicious individuals and more as a property of the architecture of cognition and known mechanisms of social learning and social relations.

The Nature of Implicit Prejudice

The discovery that prejudice can operate unwittingly, unintentionally, and unavoidably emerged from several related developments in psychology, sociology, economics, and political science. Most politically salient was the persistence of social, economic, and health-related racial discrimination despite an increasing unwillingness, during the late-twentieth century, of Americans to consciously endorse "explicit" racist attitudes (e.g., Bobo, 2001; Dovidio, 2001; Sniderman and Carmines, 1997). Although the observation of dissociations between explicit intergroup attitudes and intergroup discrimination was hardly unprecedented (e.g., Allport, 1958; La Pierre, 1934), it was met with an increasing interest in assessing political attitudes unobtrusively, either to circumvent the role of social desirability in attitude expression (e.g., Crosby, Bromley, and Saxe, 1980; Fazio et al., 1995; Word, Zanna, and Cooper, 1974), or to address the possibility that the psychology of prejudice in the United States had evolved into more sublimated, symbolic, or otherwise less deliberately hostile forms (e.g., Dovidio and Gaertner, 2004; Jackman, 1994; Sears and Henry, 2005). Equally important, developments within the information-processing paradigm of psychology made the study of implicit cognition—including automatic, implicit prejudice—both newly possible and theoretically coherent (e.g., Banaji and Greenwald, 1994; Bargh, 1999; Greenwald and Banaji, 1995). Finally, the social-psychological interest in implicit prejudice resonated with a broader interdisciplinary appreciation across the brain sciences of the variety, sophistication, and richness of information processing that occurs outside the window of conscious deliberation, indicating, among many other things, that prejudice is hardly the only kind of thinking largely implicit in nature (e.g., French and Cleeremans, 2002).

The Discovery of Implicit Prejudice

The discovery and identification of implicit prejudice as consequential, ubiquitous, and distinct from "explicit," or conscious, endorsement of prejudiced attitudes has now been firmly established by decades of research, hundreds of studies, thousands of participants from around the world, and a variety of research methodologies. Implicit prejudice was captured initially in two basic experimental paradigms that emerged from the information-processing nexus of cognitive and social psychology—one demonstrating the effects of concepts made implicitly salient through experimental manipulation, and the other demonstrating the existence and correlates of implicit semantic associations.

The effects of cognitively salient concepts on social judgment were initially captured in now-classic experiments demonstrating that evaluations of social targets are implicitly influenced by recent exposure to judgment-related information (Higgins, Rholes, and Jones, 1977; Srull and Wyer, 1979). Although interdisciplinary consensus about the importance of implicit cognition exhibited by this research tradition had been building for many years, its application to stereotyping was captured in Patricia Devine's iconic paper (1989), which marked the beginning of a

paradigm shift in the social-psychological understanding of stereotyping and prejudice more generally.[1]

In the critical experiment, participants evaluated a hypothetical person named "Donald" as more hostile if they had been subliminally exposed to a large versus a small proportion of words related to common U.S. stereotypes of African Americans. The finding was striking because it suggested that crude stereotypes could operate unintentionally and outside conscious awareness to influence social judgment, and it was disturbing because it showed that implicit stereotyping occurred to an equal degree whether participants explicitly endorsed racist attitudes or not.

This basic paradigm has since been used in scores of experiments that confirm the implicit operation of prejudice and stereotyping in social judgment including, but not limited to, ethnicity and race (e.g., Dovidio et al., 1997), gender (e.g., Rudman and Borgida, 1995), and age (e.g., Levy, 1996). As an example of the existence of implicit gender stereotypes, women but not men were judged as more dependent after recent exposure to female stereotypes, and men but not women were judged as more aggressive after exposure to male stereotypes (Banaji, Hardin, and Rothman, 1993). The effects of stereotype salience were equally large for women and men, regardless of the levels of explicit prejudice. In sum, research in this tradition suggests that mere knowledge of a stereotype can influence social judgment regardless of explicit intentions and regardless of the social category of the one doing the stereotyping.

Research demonstrating the implicit influence of cognitively salient stereotypes in social judgment has been complemented by research in the second paradigm that establishes the extent to which stereotyping and prejudice operate as webs of cognitive associations. Like Freud's discovery that mental architecture is revealed by quantifying what most easily comes to mind given targeted conceptual probes, the notion was initially captured in now-classic experiments showing that judgments on "target" words are faster if they are immediately preceded by brief exposure to semantically related, as opposed to unrelated, "prime" words (e.g., Meyer and Schvaneveldt, 1971; Neely, 1976, 1977). These semantic relations are now known to be highly correlated with those identified in free-association tasks (for a review see Ratcliff and McKoon, 1994). Extensive research demonstrates that a variety of social beliefs and attitudes function as semantic and evaluative associations across several procedural variations, including conditions in which the prime words are exposed too quickly for people to see (for reviews see Fazio, 2001; Greenwald and Banaji, 1995). For example, simple judgments about target female pronouns were faster after brief exposure to prime words either denotatively or connotatively related to women (e.g., lady, nurse) than words related to men (e.g., gentleman, doctor), and judgments about male pronouns were faster after exposure to prime words related to men than women (Banaji and Hardin, 1996; Blair and Banaji, 1996). Similarly, people were faster to judge words associated with negative stereotypes of African Americans after exposure to black faces than to white faces (e.g., Dovidio, Evans, and Tyler, 1986; Dovidio et al., 1997; Wittenbrink, Judd, and Park, 1997). Such results have been taken to demonstrate the automatic nature of beliefs or stereotypes when they capture associations between social groups and their common stereotypes, and have been used to demonstrate the automatic nature of attitudes or preferences when they capture associations between social groups and common evaluations of them.

Research in this tradition suggests the ubiquity with which common prejudice and stereotyping operates among all kinds of people along lines laid down by extant social relations on a variety of dimensions. These include, but are not limited to, ethnicity and race (e.g., Nosek, Banaji, and Greenwald, 2002a), gender (e.g., Banaji and Hardin, 1996), sexual orientation (e.g., Dasgupta and Rivera, 2008), body shape (e.g., Bessenoff and Sherman, 2000), the elderly (Perdue and Gurtman, 1990), and adolescents (Gross and Hardin, 2007). Implicit prejudice of this kind develops early in children across cultures (e.g., Baron and Banaji, 2006; Dunham, Baron, and Banaji, 2006, 2007) and appears to involve specific brain structures associated with nonrational thought (e.g., Cunningham, Nezlek, and Banaji, 2004; Lieberman, 2000; Phelps et al., 2000).

Characteristics of Implicit Prejudice

Although the identification of the course, consequences, and nature of implicit prejudice continues to evolve in research spanning disciplines, research methodologies, and specific social categories, its fundamental characteristics are now firmly established. Implicit prejudice (a) operates unintentionally and outside awareness, (b) is empirically distinct from explicit prejudice, and (c) uniquely predicts consequential social judgment and behavior. Underlying all claims about the operation of implicit prejudice is the fact that the implicit operation of stereotypes and prejudice is robust and reliably measured, as indicated by hundreds of published experiments (e.g., Banaji, 2001; Greenwald and Banaji, 1995). In addition, research shows that implicit prejudice is subject to social influence, a finding that is important to public policy considerations, although the immediate operation of

implicit prejudice is difficult, if not impossible, to control through individual volition.

The most important characteristic of implicit prejudice is that it operates ubiquitously in the course of normal workaday information processing, often outside of individual awareness, in the absence of personal animus, and generally despite individual equanimity and deliberate attempts to avoid prejudice (for reviews see Devine, 2005; Dovidio and Gaertner, 2004). Evidence of this process includes experiments demonstrating that social judgment and behavior is affected in stereotype-consistent ways by unobtrusive, and even subliminal, manipulations of stereotype salience. Typically in these kinds of experiments, participants attempt to be fair and unbiased and, moreover, exhibit no evidence of knowing that their recent experience included exposure to stereotypes used in their evaluations. Experiments that manipulate stereotype salience subliminally through extremely rapid exposure to words or images make the case especially strongly (for reviews see Bargh, 1999; Devine and Monteith, 1999). Interestingly, implicit prejudice of this kind appears to operate regardless of the personal characteristics of research participants, including participant social category, and regardless of individual differences in related explicit attitudes and implicit attitudes. The implication is that anyone who is aware of a common stereotype is likely to use it when it is cognitively salient and relevant to the judgment at hand (e.g., Hardin and Rothman, 1997; Higgins, 1996).

Complementary evidence that prejudice operates implicitly comes from research using measures of automatic cognitive association, including serial semantic priming paradigms (e.g., Blair and Banaji, 1996), subliminal serial priming paradigms (e.g., Fazio et al., 1995), and startle responses (e.g., Amodio, Harmon-Jones, and Devine, 2003), as well as behavioral interference paradigms like Stroop tasks (e.g., Bargh and Pratto, 1986; Richeson and Trawalter, 2005) and implicit association tasks (IAT; e.g., Greenwald, McGhee, and Schwartz, 1998). Hundreds of experiments using these measures suggest that people are generally surprised to learn that they have implicit prejudices.

A second major characteristic of implicit prejudice is that it is difficult for individuals to deliberately modulate, control, or fake (for reviews see Devine and Monteith, 1999; Dovidio, Kawakami, and Gaertner, 2002; Greenwald et al., 2009). Experiments like Devine's (1989), which demonstrate implicit prejudice through subliminal, unconscious manipulations of stereotype salience, by design preclude individual awareness and control, thereby demonstrating that immediate conscious awareness of stereotyped information is formally unnecessary to produce implicit stereotyping. Similar experiments that manipulate stereotype salience through recent conscious exposure to stereotyped information suggest that implicit stereotyping can occur through the kind of mere exposure to stereotyped information that occurs in the hurly-burly of everyday life in societies that are organized around race, class, and gender (e.g., Rudman and Borgida, 1995). Moreover, research expressly designed to test the success of individuals to control or fake their levels of implicit prejudice as assessed by measures of association show that it is extremely difficult or impossible to do so (Bielby, 2000), whether attitudes are about gays (e.g., Banse, Seise, and Zerbes, 2001), ethnic groups (e.g., Kim, 2003), or gender (e.g., Blair and Banaji, 1996).

Independent of individual attempts to control the operation of implicit prejudice, research shows that it is nearly impossible to consciously correct for effects of implicit prejudice (for one review see Wegener and Petty, 1997). To do so, one must be in the unlikely circumstance of having all at once (a) knowledge that implicit prejudice is operating, (b) both the motivation and cognitive capacity to control it, and perhaps most unlikely of all, (c) precise knowledge of the magnitude and direction of the correction needed (e.g., Bargh, 1999; Fazio and Towles-Schwen, 1999). For example, although individual differences in explicit prejudice predict the overt interpersonal friendliness of whites toward blacks, it is individual differences in implicit prejudice that predicts the nonverbal behavior of whites, which is the behavior that, in turn, predicts black attitudes toward whites (e.g., Dovidio, Kawakami, and Gaertner, 2002).

The third critical characteristic of implicit prejudice is that it is empirically distinct from explicit prejudice, including activating distinctive regions of the brain (Cunningham, Nezlek, and Banaji, 2004). Although explicit attitudes are often uncorrelated with the implicit operation of prejudice (e.g., Devine, 1989; Fazio and Olson, 2003) and implicit prejudiced associations (e.g., Gross and Hardin, 2007), correlations between implicit and explicit attitudes actually vary widely across studies (e.g., Hofmann et al., 2005; Nosek, 2005). A picture of when and why implicit and explicit attitudes are likely to be dissociated has begun to emerge. Baldly explicit prejudice on the basis of race and gender often conflicts with social norms of equity and justice and hence is a domain in which implicit-explicit attitude dissociations often occur. In contrast, in domains in which explicit attitudes do not conflict with consensual social norms, implicit and explicit attitudes are often correlated (e.g., Gawronski, 2002; Greenwald et al., 2009). For example, implicit prejudice is correlated with amygdala activation (Cunningham, Nezlek, and Banaji,

2004; Phelps et al., 2000), and explicit prejudice is more strongly correlated with prefrontal cortex activation (Cunningham et al., 2004; see also Amodio et al., 2004). Most importantly, implicit prejudice uniquely predicts related attitudes and behavior over and above explicit prejudice and appears to be related to distinct families of social judgment and behavior. Implicit attitudes are associated relatively more with tacit learning, manipulations, and consequences, whereas explicit attitudes are relatively more associated with intentionally controllable behaviors and attitudes (e.g., Olson and Fazio, 2003; Spalding and Hardin, 1999).

Because the unique predictive validity of implicit prejudice is critical to appreciating its implications for policy choices, we now turn to a detailed discussion of this evidence in the context of policy implications.

Consequences and Social Control of Implicit Prejudice

The existence of implicit prejudice would be of little practical consequence if it were an unreliable predictor of social judgment and behavior, particularly given the growing interest in its potential economic, labor, legal, and policy implications (e.g., Ayres, 2001; Banaji and Bhaskar, 2000; Banaji and Dasgupta, 1998; Chugh, 2004; Greenwald and Krieger, 2006; Jost et al., 2009; Kang and Banaji, 2006; Tetlock and Mitchell, in press). However, research demonstrates the consequential nature of implicit prejudice in a variety of domains, including health, job satisfaction, voting behavior, and social interaction. Our discussion of this evidence is organized around the two paradigms that led to the discovery of implicit prejudice in the first place—the implicit effects of cognitively salient stereotypes and prejudice, and the predictive utility of implicit associations between social groups and their presumed characteristics.

Implicit Effects of Cognitively Accessible Stereotypes and Prejudice

Perhaps the most disturbing aspect of implicit prejudice is that while cognitively salient stereotypes and prejudices operate outside of conscious awareness, they produce qualitative changes in social judgment and behavior. Across some two dozen experiments in which participants are presented with a series of images of social situations and instructed to as quickly and accurately as possible "shoot" if the target is armed and "don't shoot" if the target is unarmed, the finding is consistent: participants faster and more accurately shoot gun-toting black targets than white targets and faster and more accurately avoid shooting tool-toting white targets than black targets (e.g., Correll et al., 2002; Correll, Urland, and Ito, 2006). The finding is obtained among both white and black participants alike, and even among professional police officers (Correll et al., 2007; Plant and Peruche, 2005; Plant, Peruche, and Butz, 2005). In a similar experimental paradigm in which participants were instructed to distinguish between weapons and hand tools, participants were faster to correctly identify weapons after exposure to black faces than to white faces but faster to correctly identify tools after exposure to white faces than to black faces (Payne, 2001). A follow-up study demonstrated that participants under time pressure were more likely to misidentify tools as guns after exposure to black faces but misidentify guns as tools after exposure to white faces (see also Govorun and Payne, 2006; Payne, Shimizu, and Jacoby, 2005), a finding that is obtained even among professional police officers (Eberhardt et al., 2004).

Such findings have important implications for police officers, given the broader finding that police consistently use greater lethal and nonlethal force against nonwhite suspects than white suspects (e.g., for reviews see U.S. Department of Justice, 2001; Geller, 1982). Indeed, Los Angeles police officers judge adolescents accused of shoplifting or assault more negatively and as more culpable when they have been subliminally exposed to words related to common stereotypes about blacks than words that are not related to the stereotypes (Graham and Lowery, 2004).

The implicit use of common stereotypes is not limited to issues of race but is also seen in matters of age and in instances of gender bias. For example, the behavior of a seventeen-year-old (but not a seventy-one-year-old) toward a police officer is judged as more rebellious after the latter's subliminal exposure to words related to common adolescent stereotypes than with exposure to words that are not, and the magnitude of the effect is unrelated to individual differences in explicit attitudes about adolescents (Gross and Hardin, 2007). And, in a telling experiment involving stereotypes commonly traded in mass media (e.g., beer ads featuring bikini-clad models), recent exposure to sexist versus nonsexist television advertisements was shown to cause men to (a) evaluate a job applicant as more incapable and unintelligent, (b) evaluate her as more sexually attractive and receptive, (c) make more sexual advances to her, and (d) evaluate her as more deserving of being hired (Rudman and Borgida, 1995). Here, too, typical of experiments of this type, the effect of exposure to sexist ads was unqualified by individual differences in explicit endorsement of sexist beliefs and attitudes.

Implicit prejudice and stereotyping is not limited to judgments of others, however, but also affects self-judgment and behavior, especially with regard to intellectual performance. For example, Asian American women believe they are relatively better at math than verbal skills when they have identified their ethnicity, but better at verbal than math skills when they have identified their gender (e.g., Sinclair, Hardin, and Lowery, 2006). Even more striking are findings that similar manipulations implicitly affect stereotype-related intellectual performance. Consistent with the respective stereotypes, blacks, but not whites, perform worse on GRE advanced exams when ethnicity is salient (e.g., Steele and Aronson, 1995), and women, but not men, perform worse on GRE quantitative exams (Spencer, Steele, and Quinn, 1999), and worse on a logic task but not an identical verbal task, when gender is salient (Cheung and Hardin, 2010). Similarly, older, but not younger people, perform worse on memory tasks when age is salient (e.g., Levy, 1996), and students from low, but not high, socioeconomic backgrounds perform worse on intellectual tasks when economic status is salient (e.g., Croizet and Claire, 1998; Harrison et al., 2006). Moreover, gender and ethnic stereotypes can interact to produce especially large decrements in the math and spatial performance of Latina women (e.g., Gonzales, Blanton, and Williams, 2002). Such performance discrepancies are also evident via functional magnetic resonance imaging (fMRI) data. For example, women not only perform worse on mental rotation tasks when negative stereotypes are salient but performance decrements are correlated with greater activity in brain regions associated with emotion and implicit prejudice (Wraga et al., 2007).

Congruent with evidence discussed throughout this paper, the consequences of implicit prejudice to the self echo the principled operation of implicit prejudice more generally. Stereotypes are double-edged swords and hence can sometimes boost performance. For example, Asian American women perform better on quantitative tests when their ethnicity is salient than when their gender is salient (e.g., Shih, Pittinsky, and Ambady, 1999). Whether positive or negative, implicit stereotype threat effects emerge early in development and appear with increasing strength throughout elementary and middle school (e.g., Ambady et al., 2001). Finally, evidence suggests that these kinds of effects are more likely to occur when the relevant stereotypes are made salient in subtle ways rather than blatantly (Shih et al., 2002), congruent with our broader argument about the insidious role that implicit prejudice plays in everyday social cognition and behavior.

Implicit Prejudice as Cognitive Associations

Common stereotypes and prejudice not only affect social judgment and behavior implicitly, but several measures of implicit attitudes have been developed (for reviews see Olson and Fazio, 2003; Wittenbrink and Schwartz, 2007), and research based on hundreds of studies shows that implicit attitude measures are stable over time, internally consistent, and reliably predict related judgments and behaviors, including political attitudes, voting, academic achievement scores, consumer preferences, social evaluation, hiring decisions, and verbal and nonverbal affiliation (for reviews see Fazio and Olson, 2003; Nosek, 1995; Perugini, 2005). According to a recent meta-analysis (Greenwald et al., 2009), although implicit and explicit attitudes are commonly uncorrelated with each other, implicit measures are, on average, comparably correlated with criterion measures and usually more strongly correlated with measures of socially sensitive behavior than explicit measures. In short, where stereotyping and prejudice are concerned, implicit measures generally predict behavior better than explicit measures.

Unlike explicit measures, in which predictive validity often declines substantially for socially sensitive criteria, the predictive validity of implicit measures typically does not. For example, in a study reported by Rudman and Ashmore (2007), implicit prejudice uniquely predicts self-reported hostile behavior among whites toward blacks, including ethnic slurs, ostracism, and verbal and physical abuse, and does so over and above explicit attitudes and prejudice. In a second study, implicit prejudice among whites toward Jews, Asians, and blacks was shown to predict preferences to de-fund campus organizations representing Jews, Asians, and blacks, respectively—again, over and above explicit attitudes and prejudice. Implicit prejudice can also predict prejudice-related judgments when explicit attitudes do not, particularly in cases of intergroup relations (reviewed in Greenwald et al., 2009). For example, unlike explicit prejudice, implicit racial prejudice among whites predicts quickness to perceive anger in black faces but not white faces (Hugenberg and Bodenhausen, 2003).

It is one thing for individual differences in implicit prejudice to predict attitudes and judgment, but it is quite another for it to predict behavior. Implicit attitudes predict nonverbal friendliness and discomfort of whites when interacting with blacks (Dovidio et al., 1997, 2002) and how positively blacks perceive whites with whom they interact (Dovidio, Kawakami, and Gaertner, 2002; Fazio et al., 1995; Sekaquaptewa et al., 2003). For example, in research particularly

telling for common educational and school situations, Richeson and Shelton (2005) found that in face-to-face interpersonal interactions, individual differences in implicit prejudice were more apparent to black than white perceivers and more apparent when whites interacted with blacks than with other whites (see also Perugini, O'Gorman, and Prestwich, 2007; Ziegert and Hanges, 2005).

Implicit attitudes not only affect social judgment and behavior relative to others but also are important predictors of one's own behavior and self-evaluation. For example, implicit, but not explicit, self-esteem predicts anxious behavior in self-threatening situations but not in unthreatening situations (Spalding and Hardin, 1999; see also Asendorpf, Banse, and Mucke, 2002; Egloff and Schmukle, 2002). Women who implicitly associate romance with chivalry report less interest in economic and educational achievement (Rudman and Heppen, 2003), and implicit dissociations between the concepts of math and women predict lower quantitative SAT scores among women (Nosek, Banaji, and Greenwald, 2002b). Finally, a surprising number of African Americans exhibit implicit preference for whites over blacks (e.g., Nosek, Banaji, and Greenwald, 2002a). Variability in implicit antiblack prejudice among African Americans predicts stated preferences for working with white versus black partners on intellectually demanding tasks and does so independently of explicit attitudes (Ashburn-Nardo, Knowles, and Monteith, 2003), a finding suggesting that the general tendency to favor in-groups over out-groups may be trumped by implicit stereotypes relevant to the task at hand (see also Rudman, Feinberg, and Fairchild, 2002).

Most of the research on the predictive validity of implicit prejudice discussed thus far involves undergraduate participant samples in laboratory settings, yet one might rightly wonder whether implicit prejudice will matter in daily tasks, big and small. One reason to believe that it will is research showing that among people who have finished their formal education, implicit attitudes predict behavior and judgment on dimensions that matter to people beside college students and do so on a variety of dimensions of undeniable real-world application. For example, implicit attitudes predict suicide attempts (Glashouwer et al., 2010; Nock and Banaji, 2007; Nock et al., 2010), severity and treatment outcomes for phobia and panic disorders (e.g., Teachman, Marker, and Smith-Janik, 2008; Teachman, Smith-Janik, and Saporito, 2007; Teachman and Woody, 2003), condom use (Marsh, Johnson, and Scott-Sheldon, 2001), smoking status (Swanson, Rudman, and Greenwald, 2001), alcohol consumption (Weirs et al., 2002), and consumer preferences for consumer goods like yogurt, beverages, and fast-food restaurants (Maison, Greenwald, and Bruin, 2004). In addition, reductions in implicit romantic attraction predict the subsequent breakup of committed relationships (Lee, Rogge, and Reis, 2010).

In addition to the large and growing literature demonstrating the predictive validity of measures of implicit attitudes in matters of everyday life, research shows that implicit prejudice predicts behavior outside the laboratory. For example, implicit preference among Swedish job recruiters for native Swedes over Arabs predicts interview preferences (Rooth, 2010). Overall, native Swedes were more than three times more likely to receive interview callbacks than equally qualified Arabs.

Several studies demonstrate that implicit prejudice predicts voting behavior, including the historic 2008 election in which Barack Obama became the first African American to be elected president of the United States. For example, in the week before the election, implicit antiblack prejudice predicted intention to vote for John McCain over Obama and did so independently of self-reported conservatism (Greenwald et al., 2009). Another study found that the degree to which participants implicitly associated America more with McCain than Obama predicted intention to vote for McCain (Devos and Ma, 2010).

Implicit prejudice not only predicts voting intentions before elections but also reported voting behavior after elections. Voters were substantially less likely to report voting for Barack Obama, and exhibited more negative attitudes toward health care reform, the greater their implicit prejudice (Knowles, Lowery, and Shaumberg, 2010), and, in a follow-up study conducted nearly a year after the election, implicit prejudice remained a significant predictor of negative attitudes toward Obama. Moreover, implicit prejudice predicted negative attitudes about health-care reform when it was ascribed to Obama but not when the identical reform was ascribed to Bill Clinton. Similar findings have obtained in studies of the Italian electorate, as well (e.g., Arcuri et al., 2008; Galdi, Arcuri, and Gawronski, 2008; Roccato and Zogmaister, 2010).

Another area of society in which the real-world operation of implicit prejudice is implicated is in the practice of medicine, in which differential treatment as a function of ethnicity is a well-documented case in point. A recent study of emergency-room treatment of more than 150,000 patients complaining of severe pain over a 13-year span found that whites were given powerful opioid pain killers more than blacks and Hispanics, with evidence suggesting that

the disparity is due more to undertreatment of minorities rather than overtreatment of whites (Pletcher et al., 2008). Racial disparities are well documented for treatment of cardiovascular disease as well (for a review see Kressin and Petersen, 2001), including expensive treatments for acute myocardial infarction (e.g., Petersen et al., 2002).

New evidence suggests that at least one cause for such findings may be individual differences in implicit prejudice among treating physicians. In a study that assessed both explicit and implicit attitudes toward whites and blacks and treatment recommendations for hypothetical patients who differed only as a function of an experimental manipulation of race, emergency-room physicians exhibited strong implicit preference for whites over blacks, and also strong implicit associations of blacks versus whites for being uncooperative, despite exhibiting no explicit preferences for whites or differences in cooperativeness between whites and blacks. Importantly, however, although explicit attitudes did not predict emergency treatment recommendations, implicit attitudes did. Greater implicit prejudice predicted an increasing likelihood to treat whites and a decreasing likelihood to treat blacks exhibiting identical symptoms (Green et al., 2007). By extension, and perhaps unsurprisingly, implicit racial bias among physicians negatively predicts African American patient satisfaction with their physicians (Penner et al., 2010).

Consistent with laboratory findings suggesting that implicit attitudes should be uniquely strong predictors of counternormative behavior, implicit negative attitudes toward injection-drug users among drug and alcohol nurses who treat them predicts nurses' stated intentions to leave drug and alcohol nursing, over and above relevant explicit attitudes (von Hippel, Brener, and von Hippel, 2008),[2] corroborating laboratory demonstrations of the unique predictive power of implicit measures when judgments are potentially nonnormative (Greenwald et al., 2009). In other words, although the medical model frames drug and alcohol abuse as an involuntary disease to be treated, and as such abusers should be worthy of sympathy, the day-to-day experience with a population known to be difficult and challenging by a part of the medical community that is known to have a high job turnover rate may make expressly negative attitudes about abusers counternormative. In addition, it is implicit prejudice (but not explicit prejudice) that mediates the well-documented relation between stress and intention to change jobs (von Hippel, Brener, and von Hippel, 2008).

In short, research demonstrating the real-world applicability of implicit attitudes continues to grow, and it is no longer credible to hide behind the view that the predictive validity of implicit prejudice on judgment and behavior is a quirk of the laboratory (see also Jost et al., 2009).

Social Control of Implicit Prejudice

Given evidence that implicit prejudice is reliably captured and measured and that it is consequential, ubiquitous, and stubbornly immune to individual attempts to control it, what hope is there for effective policy solutions? Although implicit prejudice presents challenges to public policy formulations based on outdated notions of the nature of prejudice, recent research shows that it behaves in predictable ways that conform to fundamental principles of social and cognitive psychology. Implicit prejudice reflects stable social relationships and organization by reflecting social identities, group categorizations and status, as well as general preferences for the self, similar others, and ingroups (e.g., Bosson, Swann, and Pennebaker, 2000; Greenwald, McGhee, and Schwartz, 1998; Spalding and Hardin, 1999). Moreover, evidence suggests that implicit prejudice is responsive to social dynamics, including (a) relative intergroup status (e.g., Rudman, Feinberg, and Fairchild, 2002), (b) minimal group categorization (Ashburn-Nardo, Voils, and Monteith, 2001), (c) chronic and temporary changes in the salience of prejudice-related information (e.g., Dasgupta and Greenwald, 2001), and (d) friendly intergroup contact (e.g., Tam et al., 2006). Implicit prejudice can also increase and decrease as a function of conditioning that is consistent with the fundamentals of learning theory (e.g., Bargh, 1996; Fazio 2001, 2003; Fazio and Olson, 2003; Hardin and Rothman, 1997), and it generally conforms to principles of cognitive consistency (e.g., Greenwald et al., 2009).

An obvious but important indication of the way implicit prejudice reflects social dynamics is the fact that it so well tracks the character of chronic social organization, including relative group power, social status, and concomitant stereotypes. For example, although in-group preference is a common feature of implicit prejudice (e.g., Greenwald et al., 1998), at least as important are findings that it reflects social status. Members of high-status groups in the United States not only exhibit greater implicit group favoritism than low-status groups but also do so as a function of their relative status, whether they are rich, white, skinny, or Christian (e.g., Nosek et al., 2002a; Rudman, Feinberg, and Fairchild, 2002). However, at the same time, although in-group preference is common in both implicit and explicit prejudice, outgroup preference is hardly rare (e.g., Jost and Banaji, 1994) and also closely aligns with relative group

status. For example, members of low-status groups were more likely to implicitly favor dominant out-groups to the extent that their in-group was low in status, despite exhibiting strong explicit in-group favoritism (Jost, Pelham, and Carvallo, 2002; Rudman, Feinberg, and Fairchild, 2002).

Implicit prejudice not only reflects stable social and organizational hierarchies, but research shows that changes in social organization also predict corresponding changes in implicit prejudice, a finding that has promising implications for public policy. Friendly intergroup contact is shown to reduce both implicit and explicit prejudice alike (e.g., Henry and Hardin, 2006; Turner, Hewstone, and Voci, 2007). In one example, implicit prejudice toward gay and lesbian people was found to be lower for people who reported high levels of long-term contact with gay and lesbian people as well as for people who reported being exposed to gay-positive media (Cheung et al., 2011; Dasgupta and Rivera, 2008). Similarly, implicit prejudice toward the elderly was lower among college students the more friendships they reported having with older people (Tam et al., 2006). In yet another example, implicit prejudice was found to be lower between British and South Asian children in England to the extent that they reported out-group friendships, and implicit prejudice was reduced even among children who reported no out-group friendships themselves but who reported having friends who did (Turner, Hewstone, and Voci, 2007). Causal modeling in this research indicates that the findings are more consistent with intergroup friendships affecting implicit prejudice than with implicit prejudice affecting friendship patterns (Tam et al., 2006; Turner, Hewstone, and Voci, 2007), a conclusion corroborated experimentally. For example, implicit prejudice among white college freshmen was reduced more over the course of their first school term if they were randomly assigned to a black roommate than a white roommate (Shook and Fazio, 2007).

Although friendly intergroup contact generally reduces implicit intergroup prejudice, recent findings demonstrate that intergroup contact does not always have purely positive outcomes. For example, anti-adolescent implicit prejudice among adolescents was greater to the degree that they reported having close friendships with adults (Gross and Hardin, 2007). Evidence also suggests that relatively stable aspects of social hierarchy complicate matters. In research involving blacks and whites in Chicago and Christians and Muslims in Lebanon, implicit intergroup prejudice was shown to be lower to the degree that participants reported out-group friendships (Henry and Hardin, 2006). However, results also indicate that implicit prejudice reduction is greater for low-status group members toward high-status group members than it is for high-status group members toward low-status group members. That is, in this study, out-group friendships predicted greater reductions in implicit prejudice for Muslims than Christians and for blacks than whites due to their places in the social hierarchy.

Research also indicates that implicit prejudice is affected by social dynamics throughout development (e.g., Baron and Banaji, 2006; Rutland et al., 2005) and that the development of implicit prejudice is likely to be bound up with interpersonal dynamics involving interpersonal identification and intersubjectivity (e.g., Hardin and Conley, 2001; Hardin and Higgins, 1996). For example, implicit intergroup prejudice between Korean and Japanese students in the United States was greater to the degree that participants remained connected to their ethnic heritage as indicated by linguistic fluency (Greenwald, McGhee, and Schwartz, 1998). People exhibited more positive implicit attitudes toward women to the degree that they reported being raised more by their mothers than their fathers (Rudman and Goodwin, 2004). And, implicit racial prejudice among white fourth- and fifth-grade children was correlated with the explicit prejudice of their parents, but only to the extent that they identified with their parents (Sinclair, Lowery, and Dunn, 2005), and the implicit prejudice of mothers predicted racial preferences exhibited by their three- to six-year-old children (Castelli, Zogmaister, and Tomelleri, 2009).

Research demonstrating the long-term social determinants of implicit prejudice is likely to be either encouraging or depressing, depending upon one's sense of the likelihood of broad, long-term changes in social organization and culture. It is important, however, to remember that such things do happen. What changes in implicit prejudice might be revealed if the measures had been in existence long enough to reflect suffrage, women's mass entry into the workforce during World War II, the civil rights movement, and twentieth-century urban white flight, to name just a few societal sea changes?

Although we believe that culture-wide changes in implicit prejudice will require culture-wide changes in social organization and practice, another way in which implicit prejudice obeys principles of social psychology offers some promise of more immediate, if local, opportunities for progress. Research shows that implicit prejudice is subject to the demands of immediate situations and interpersonal dynamics, much like human behavior more generally (e.g., Ross and Nisbett, 1991). For example, white participants exhibited lower implicit prejudice in the presence of a black experimenter than a white experimenter

(Lowery, Hardin, and Sinclair, 2001; Richeson and Ambady, 2003). Interestingly, however, Lowery and colleagues (2001) also found that this automatic social tuning effect did not occur among Asian American participants, whose implicit prejudice was reduced only when the experimenter expressly told them to avoid prejudice. This finding suggests that although the norm to avoid prejudice may operate tacitly for some, it may require explication for people who do not yet recognize their potential role as ciphers of prejudice.

Research also suggests that the interpersonal regulation of implicit prejudice is due in part to a motivation to affiliate with others who are presumed to hold specific values related to prejudice, as implied by shared reality theory (e.g., Hardin and Conley, 2001). For example, participants exhibited less implicit racial prejudice in the presence of an experimenter wearing a T-shirt with an antiracism message than a blank T-shirt, but only when the experimenter was likeable (Sinclair et al., 2005). When the experimenter was not likeable, implicit prejudice was actually greater in the presence of the ostensibly egalitarian experimenter. In addition, social tuning in these experiments was mediated by the degree to which participants liked the experimenter, providing converging evidence that interpersonal dynamics play a role in the modulation of implicit prejudice, as they do in other dimensions of social cognition (Hardin and Conley, 2001; Hardin and Higgins, 1996).

As regards public and personal policy, these findings suggest that a public stance for egalitarian values is a double-edged sword, and a sharp one at that. Although it may reduce implicit prejudice among others when espoused by someone who is likeable and high in status, it may backfire when espoused by someone who is not likeable or otherwise of marginal status. This finding suggests one mechanism by which common forms of "sensitivity training" in service of the reduction of workplace sexism and racism may be subverted by interpersonal dynamics, however laudable the goals.

Demonstrating the utility of specific interventions to reduce implicit prejudice, Rudman, Ashmore, and Gary (2001) found that diversity education with a likeable black professor reduced implicit prejudice and did so through liking for the professor, increased friendships with other African Americans, and reduced fear of blacks. Likewise, thinking about gay-positive role models reduced implicit prejudice for those with low contact with gay and lesbian people to the level of those with high contact and increased the endorsement of gay-positive attitudes, including legalizing civil unions for gays and lesbians (Dasgupta and Rivera, 2008).

In a cautionary note, however, the lack of long-term exposure to a particular group can sometimes trigger greater implicit prejudice when a member of the group is present. In one example, people who reported having no gay friends at all exhibited greater implicit antigay prejudice when a male experimenter incidentally mentioned his "boyfriend" than when he mentioned his "girlfriend." Similarly, women who reported having no lesbian friends exhibited greater implicit antilesbian bias when the experimenter was from a gay and lesbian organization (Cheung et al., 2011). This research complements research showing immediate social influence on implicit prejudice. It suggests that as powerful as immediate social norms might be, implicit prejudice is ultimately expressed differently from individual to individual as a function of attitudes presumed to be held by others in relevant long-term social relationships, sometimes in subtle or even contradictory ways, much as it depends on other dimensions of social cognition (e.g., Hardin and Higgins, 1996).

Research demonstrating that implicit prejudice is subject to social influence is broadly consistent with principles of information processing (for a review see Blair, 2002). Implicit racial prejudice is reduced (a) when admired black exemplars are used (e.g., Dasgupta and Greenwald, 2001; cf. De Houwer, 2001), (b) after seeing an image of blacks at a friendly barbeque versus unfriendly street corner (Wittenbrink, Judd, and Park, 2001), and (c) imagining the virtues of multicultural education (Richeson and Nussbaum, 2004). In contrast, implicit racial prejudice is increased after exposure to violent rap music (Rudman and Lee, 2002). Implicit gender stereotyping is reduced for those who have recently been exposed to images of female leaders (Dasgupta and Asgari, 2004) or have recently imagined a powerful woman (Blair, Ma, and Lenton, 2001). This research suggests that simple images and text in immediate situations can affect levels of implicit prejudice for those in the situation in ways that are broadly congruent with construct accessibility theory (e.g., Bargh, 1996), which is the "common language" that underlies most information-processing theory in social cognition (Higgins, 1996).

Taken together, research on the social control of implicit prejudice is broadly congruent with the Marxian maxim that egalitarian societies elicit egalitarian-minded people, as well as with the Skinnerian maxim that admirable individual behavior is elicited by situations that reinforce admirable behavior. Indeed, the methodological and theoretical advances that have transformed the understanding of the nature of prejudice—including sometimes-puzzling relations between implicit and explicit

prejudice—resonates with what Skinner argued about the relation between scientific advances and the understanding of human nature more generally:

> The line between public and private is not fixed. The boundary shifts with every discovery of a technique for making private events public . . . The problem of privacy may, therefore, eventually be solved by technical advance.
>
> —*B.F. Skinner, 1953, p.282*

Conclusions

It is not far-fetched to argue that successful policy solutions to the problem of prejudice are best pursued in light of the science of the nature of prejudice. Research in recent decades has revealed the insidious capacity of prejudice to operate implicitly—unwittingly, unintentionally, and unavoidably—as well as its course, consequences, and control at the nexus of individual cognition and social relations. In some ways, the transformative understanding of the nature of prejudice brings full circle the story of human nature since its inception in American social psychology in the mid-twentieth-century work of Sherif, Lewin, Asch, and others as an attempt to understand how seemingly good people can participate in genocide, which is also captured in Hannah Arendt's memorable phrase, "the banality of evil."

Indeed, the most important thing to know about the nature of prejudice is that it is ever present in human behavior and cognition. It remains sufficiently in the background such that it eludes conscious awareness and immediate individual control, yet it is often consequential in everyday life. Its capacity to affect social judgment and behavior without personal animus or hostility is dismissed or ignored at some peril, because a continued focus on the problem of prejudice as a result of the nonnormatively hostile behavior of the few is likely to distract policy makers from adopting strategies more strongly rooted in the science of the many. What remains are questions about how best to deal with these discoveries in shaping personal and public policy—questions that are in this light only beginning to receive the empirical attention they deserve.

What must enter into any policy computation are additional facts about the nature of prejudice beyond the primary idea that banality is its modus operandi. We must add to this the idea that prejudices and stereotypes are rooted in social consensus; they are not random. Within a given society, the likes, dislikes, and beliefs that constrain some and privilege others occur in patterns that systematically oppress subordinates while further ingraining the superiority of the dominants. Were the effects of prejudice and stereotypes less systematic, policy intervention would be less needed because their effects may be said to cancel each other out. However, when, for example, over 80% of American whites and Asians show antiblack bias and over 90% of Americans show anti-elder bias, we must pay heed. Policies that are willing to take into account the presence of implicit forms of prejudice and discrimination as a given will be the more forward-thinking instruments for change because they will be rooted in a truth about human nature and social contexts.

Furthermore, for societies that derive their sense of good character on the basis of personal accomplishment and meritocracy, research on implicit prejudice poses particularly thorny problems. The research we reviewed suggests that behavior is shaped by the social jostling and "sloshing around" of the individual, unbeknownst to the person and those around her, suggesting that the problem of implicit prejudice may be especially insidious in a society that celebrates, evaluates, and is organized around individual meritocracy. Indeed, research shows that beliefs in meritocracy pose special problems for members of stigmatized groups (e.g., Jost and Burgess, 2000; Jost and Thompson, 2000). For example, Filipina domestic workers in Hong Kong, as well as women in the United States, devalued the monetary value of their work more if their group identity was salient, but do so only to the degree that they endorsed system-justifying attitudes related to meritocracy (Cheung and Hardin, 2010). The aggregation of these kinds of effects, both large and small, but systematically organized across situations and social roles, suggests at the very least the possibility that even incrementally small biases may be expressed through actions that create a large divide among people.

Research demonstrating the effects of stereotypes and prejudice on behavior give direction to policy makers for the types of behavior most in need of their attention. It is our contention that locating the problem of prejudice in a few problematic individuals and designing solutions to the problem around this view is to miss the point. The profound implication of the discovery of implicit prejudice is that anybody is capable of prejudice, whether they know it or not, and of stereotyping, whether they want to or not. Therefore, given the implicit operation of prejudice and stereotyping and its ubiquitous nature, we believe that solutions should focus on identifying the enabling conditions that call out prejudice and stereotyping across individuals rather than focusing on identifying the rotten apples. Once identified, we must focus on the enabling conditions that promote egalitarianism

and healthy individuation. What kinds of situations bring out implicit egalitarian attitudes? Congruent with well-documented principles identified across the behavioral and mind sciences and corroborated in research on implicit prejudice, social situations populated with powerful, likeable people who are known or assumed to hold egalitarian values implicitly call out like minds in those around them.

Notes

We thank Sanden Averett, Rick Cheung, John Jost, Michael Magee, Eldar Shafir, and two anonymous reviewers for thoughtful comments on a previous draft of this paper.

1. Here and throughout we adopt conventions of social-psychological nomenclature in our use of terms. The umbrella term *attitude* includes evaluations (prejudice), beliefs (stereotypes), and behaviors (discrimination) regarding an attitude object. The terms *explicit* and *implicit* are used to capture a well-accepted heuristic dichotomy between modes of mental functions that operate largely consciously and reflectively versus unconsciously and automatically. Hence, *implicit attitude* refers to the strength of automatic association between an attitude object and characteristic attributes, *implicit prejudice* refers to the strength of automatic associations between social groups and attributes good and bad, and *implicit stereotyping* refers to the strength of automatic associations between social groups and characteristic attributes which may vary in evaluative valence.

2. Specific intention to change jobs is the strongest known predictor of actual voluntary job changes (van Breukelen, van der List, and Steensma, 2004).

References

Allport, G. W. (1958). *The nature of prejudice*. New York: Doubleday.

Ambady, N., Shih, M., Kim, A., and Pittinsky, T. L. (2001). Stereotype susceptibility in children: Effects of identity activation on quantitative performance. *Psychological Science*, *12*, 385–390.

Amodio, D. M., Harmon-Jones, E., and Devine, P. G. (2003). Individual differences in the activation and control of affective race bias as assessed by startle eyeblink response and self-report. *Journal of Personality and Social Psychology*, *84*, 738–753.

Amodio, D. M., Harmon-Jones, E., Devine, P. G., Curtin, J. J., Hartley, S. L., and Covert, A. E. (2004). Neural signals for the detection of unintentional race bias. *Psychological Science*, *15*, 88–93.

Arcuri, L., Castelli, L., Galdi, S., Zozmaister, C., and Amadori, A. (2008). Predicting the vote: Implicit attitudes as predictors of the future behavior of the decided and undecided voters. *Political Psychology*, *29*, 369–387.

Asendorpf, J. B., Banse, R., and Mucke, D. (2002). Double dissociation between implicit and explicit personality self-concept: The case of shy behavior. *Journal of Personality and Social Psychology*, *83*, 380–393.

Ashburn-Nardo, L., Knowles, M. L., and Monteith, M. J. (2003). Black Americans' implicit racial associations and their implications for intergroup judgment. *Social Cognition*, *21*, 61–87.

Ashburn-Nardo, L., Voils, C. I., and Monteith, M. J. (2001). Implicit associations as the seeds of intergroup bias: How easily do they take root? *Journal of Personality and Social Psychology*, *81*, 789–799.

Ayres, I. (2001). *Pervasive prejudice? Unconventional evidence of race and gender discrimination*. Chicago: University of Chicago Press.

Banaji, M. R. (2001). Implicit attitudes can be measured. In H. I. Roediger and J. S. Nairne (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 117–150). Washington, DC: American Psychological Association.

Banaji, M. R., Bazerman, M., and Chugh, D. (2003, December). How (un)ethical are you? *Harvard Business Review*, pp. 56–64.

Banaji, M. R., and Bhaskar, R. (2000). Implicit stereotypes and memory: The bounded rationality of social beliefs. In D. L. Schacter and E. Scarry (Eds.), *Memory, brain, and belief* (pp. 139–175). Cambridge, MA: Harvard University Press.

Banaji, M. R., and Dasgupta, N. (1998). The consciousness of social beliefs: A program of research on stereotyping and prejudice. In V. Y. Yzerbyt, G. Lories, and B. Dardenne (Eds.), *Metacognition: Cognitive and social dimensions* (pp. 157–170). Thousand Oaks, CA: Sage.

Banaji, M. R., and Greenwald, A. G. (1994). Implicit stereotyping and prejudice. In M. P. Zanna and J. M. Olson (Eds.), *The psychology of prejudice: The Ontario Symposium* (Vol. 7, pp. 55–76). Hillsdale, NJ: Lawrence Erlbaum.

Banaji, M. R., and Hardin, C. D. (1996). Automatic gender stereotyping. *Psychological Science*, *7*, 136–141.

Banaji, M. R., Hardin, C., and Rothman, A. J. (1993). Implicit stereotyping in person judgment. *Journal of Personality and Social Psychology*, *65*, 272–281.

Banse, R., Seise, J., and Zerbes, N. (2001). Implicit attitudes towards homosexuality: Reliability, validity, and controllability of the IAT. *Zeitschrift fur Experimentelle Psychologie*, *48*, 145–160.

Bargh, J. A. (1996). Principles of automaticity. In E. T. Higgins and A. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 169–183). New York: Guilford.

———. (1999). The cognitive monster: The case against the controllability of automatic stereotype effects. In S. Chaiken and Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 361–382). New York: Guilford.

Bargh, J. A., and Pratto, F. (1986). Individual construct

accessibility and perceptual selection. *Journal of Experimental Social Psychology*, *49*, 1129–1146.

Baron, A. S., and Banaji, M. R. (2006). The development of implicit attitudes: Evidence of race evaluations from ages 6, 10 and adulthood. *Psychological Science*, *17*, 53–58.

Bessenoff, G. R., and Sherman, J. W. (2000). Automatic and controlled components of prejudice toward fat people: Evaluation versus stereotype activation. *Social Cognition*, *18*, 329–353.

Bielby, W. T. (2000). Minimizing workplace gender and racial bias. *Contemporary Sociology*, *29*, 120–129.

Blair, I. V. (2002). The malleability of automatic stereotyping and prejudice. *Journal of Personality and Social Psychology Review*, *6*, 242–261.

Blair, I. V., and Banaji, M. R. (1996). Automatic and controlled processes in stereotype priming. *Journal of Personality and Social Psychology*, *70*, 1142–1163.

Blair, I. V., Ma, J. E., and Lenton, A. P. (2001). Imagining stereotypes away: The moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology*, *81*(5), 828–841. doi:10.1037/0022-3514.81.5.828

Bobo, L. (2001). Racial attitudes and relations at the close of the twentieth century. In N. Smelser, W. J. Wilson, and F. Mitchell (Eds.), *America becoming: Racial trends and their consequences* (pp. 262–299). Washington, DC: National Academy Press.

Bosson, J. K., Swann, W. and Pennebaker, J. W. (2000). Stalking the perfect measure of self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social Psychology*, *79*, 631–643.

Castelli, L., Zogmaister, C., and Tomelleri, S. (2009). The transmission of racial attitudes within the family. *Developmental Psychology*, *45*, 586–591.

Cheung, R. M., Fingerhut, A., Johnson, A., Noel, S., Drus, M., and Hardin, C. D. (2011). *Religiosity, heterosexuality, and anti-gay prejudice: Shared norms in everyday social tuning*. Unpublished manuscript, Department of Psychology, Brooklyn College, CUNY.

Cheung, R. M., and Hardin, C. D. (2010). Costs and benefits of political ideology: The case of economic self-stereotyping and stereotype threat. *Journal of Experimental Social Psychology*, *46*(5), 761–766. doi:10.1016/j.jesp.2010.03.012

Chugh, D. (2004). Societal and managerial implications of implicit social cognition: Why milliseconds matter. *Social Justice Research*, *17*, 203–222.

Correll, J., Park, B., Judd, C. M., and Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, *83*, 1314–1329.

Correll, J., Park, B., Judd, C. M., Wittenbrink, B., Sadler, M. S., and Keesee, T. (2007). Across the thin blue line: Police officers and racial bias in the decision to shoot. *Journal of Personality and Social Psychology*, *92*, 1006–1023.

Correll, J., Urland, G. L., and Ito, T. A. (2006). Event-related potentials and the decision to shoot: The role of threat perception and cognitive control. *Journal of Experimental Social Psychology*, *42*, 120–128.

Croizet, J., and Claire, T. (1998). Extending the concept of stereotype threat to social class: The intellectual underperformance of students from low socioeconomic backgrounds. *Personality and Social Psychology Bulletin*, *24*, 588–594.

Crosby, F., Bromley, S., and Saxe, L. (1980). Recent unobtrusive studies of Black and White discrimination and prejudice: A literature review. *Psychological Bulletin*, *87*, 546–563.

Cunningham, W. A., Johnson, M. K., Raye, C. L., Gatenby, J. C., Gore, J. C., and Banaji, M. R. (2004). Separable neural components in the processing of black and white faces. *Psychological Science*, *15*(12), 806–813. doi:10.1111/j.0956-7976.2004.00760.x

Cunningham, W. A., Nezlek, J. B., and Banaji, M. R. (2004). Implicit and explicit ethnocentrism: Revisiting the ideologies of prejudice. *Personality and Social Psychology Bulletin*, *30*(10), 1332 –1346. doi:10.1177/0146167204264654

Darley, J. M., and Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, *44*, 20–23.

Dasgupta, N., and Asgari, S. (2004). Seeing is believing: Exposure to counterstereotypic women leaders and its effect on automatic gender stereotyping. *Journal of Experimental Social Psychology*, *40*, 642–658.

Dasgupta, N., and Greenwald, A. G. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology*, *81*, 800–814.

Dasgupta, N., and Rivera, L. M. (2008). When social context matters: The influence of long-term contact and short-term exposure to admired outgroup members on implicit attitudes and behavioral intentions. *Social Cognition*, *26*, 54–66.

De Houwer, J. (2001). A structure and process analysis of the IAT. *Journal of Experimental Social Psychology*, *37*, 443–451.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*, 5–18.

———. (2005). Prejudice with and without compunction: Allport's inner conflict revisited. In J. F. Dovidio, P. Glick, and L. A. Rudman (Eds.), *On the nature of prejudice: Fifty years after Allport* (pp. 327– 342). Oxford: Blackwell.

Devine, P. G., and Monteith, M. J. (1999). Automaticity and control in stereotyping. In S. Chaiken and Y. Trope (Eds.), *Dual-process models and themes in social and cognitive psychology* (pp. 339–360). New York: Guilford Press.

Devos, T., and Ma, D. S. (2010). *How "American" is Barack Obama? The role of national identity in a historic bid for the White House.* Unpublished manuscript, Department of Psychology, University of Chicago.

Dovidio, J. F. (2001). On the nature of contemporary prejudice: The third wave, *Journal of Social Issues*, *57*, 829–849.

Dovidio, J. F., Evans, N. and Tyler, R. B. (1986). Racial stereotypes: The contents of their cognitive representations. *Journal of Experimental Social Psychology*, *22*, 22–37.

Dovidio, J. F., and Gaertner, S. L. (2004). Aversive racism. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 36, pp. 1–51). San Diego, CA: Academic Press.

Dovidio, J. F., Kawakami, K., and Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, *82*, 62–68.

Dovidio, J. F., Kawakami, K., Johnson, C., Johnson, B., and Howard, A. (1997). On the nature of prejudice: Automatic and controlled processes. *Journal of Experimental Social Psychology*, *33*, 510–540.

Duncan, B. L. (1976). Differential social perception and attribution of intergroup violence: Testing the lower limits of stereotyping of blacks. *Journal of Personality and Social Psychology*, *34*, 590–598.

Dunham, Y., Baron, A. S., and Banaji, M. R. (2006). From American city to Japanese village: A cross-cultural investigation of implicit race attitudes. *Child Development*, *77*, 1268–1281.

———. (2007). Children and social groups: A developmental analysis of implicit consistency among Hispanic-Americans. *Self and Identity*, *6*, 238–255.

Duraisingam, V., Pidd, K., Roche, A. M., and O'Connor, J. (2006). *Stress, satisfaction and retention among alcohol and other drug workers in Australia*. Adelaide, Australia: National Centre for Education and Training on Addiction.

Eberhardt, J. L., Goff, P. A., Purdie, V. J., and Davies, P. G. (2004). Seeing black: Race, crime, and visual processing. *Journal of Personality and Social Psychology*, *87*, 876–893.

Egloff, B., and Schmukle, S. C. (2002). Predictive validity of an Implicit Association Test for assessing anxiety. *Journal of Personality and Social Psychology*, *83*, 1441–1455.

Fazio, R. H. (2001). On the automatic activation of associated evaluations: An overview. *Cognition and Emotion*, 15, 115–141.

———. (2003). Variability in the likelihood of automatic attitude activation: Data reanalysis and commentary on Bargh, Chaiken, Govender and Pratto (1992). *Journal of Personality and Social Psychology*, *64*, 753–758.

Fazio, R. H., Jackson, J. R., Dunton, B. C., and Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, *69*, 1013–1027.

Fazio, R. H., and Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and uses. *Annual Review of Psychology*, *54*, 297–327.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., and Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, *50*, 229–238.

Fazio, R. H., and Towles-Schwen, T. (1999). The MODE model of attitude-behavior processes. In S. Chaiken and Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 97–116). New York: Guilford.

French, R. M., and Cleeremans, A. (2002). *Implicit learning and consciousness: An empirical, philosophical, and computational consensus in the making.* Hove, UK: Psychology Press.

Galdi, S., Arcuri, L., and Gawronski, B. (2008). Automatic mental associations predict future choices of undecided decision-makers. *Science*, *321*, 1100–1102.

Gallon, S. L., Gabriel, R. M., and Knudsen, J.R.W. (2003). The toughest job you'll ever love: A Pacific Northwest treatment workforce survey. *Journal of Substance Abuse Treatment*, *24*, 183–196.

Gawronski, B. (2002). What does the implicit association test measure? A test of the convergent and discriminant validity of prejudice-related IATs. *Experimental Psychology*, *49*, 171–180.

Geller, W. A. (1982). Deadly force: What we know. *Journal of Police Science and Administration*, *10*, 151–177.

Glashouwer, K. A., de Jong, P. J., Penninx, B.W.J.H., Kerkhof, A.J.F.M., van Dyck, R., and Ormel, J. (2010). Do automatic self-associations relate to suicidal ideation? *Journal of Psychopathology and Behavioral Assessment*, *32*, 428–437. doi:10.1007/s10862-009-9156-y

Gonzales, P. M., Blanton, H., and Williams, K. J. (2002). The effects of stereotype threat and double-minority status on the test performance of Latino women. *Personality and Social Psychology Bulletin*, *28*, 659–670.

Govan, C. L., and Williams, K. D. (2004). Changing the affective valence of stimulus items influences the IAT by re-defining the category labels. *Journal of Personality and Social Psychology*, *40*, 357–365.

Govorun, O., and Payne, B. K. (2006). Ego depletion and prejudice: Separating automatic and controlled components. *Social Cognition*, *24*, 111–136.

Graham, S., and Lowery, B. S. (2004). Priming unconscious racial stereotypes about adolescent offenders. *Law and Human Behavior*, *28*, 483–504.

Green, A. R., Carney, D. R., Pallin, D. J., Ngo, L. H., Raymond, K. L., Iezzoni, L. I., and Banaji, M. R. (2007).

Implicit bias among physicians and its prediction of thrombolysis decisions for black and white patients. *Journal of General Internal Medicine*, *22*, 1231–1238.

Greenwald, A. G., and Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*, 4–27.

Greenwald, A. G., and Krieger, L. H. (2006). Implicit bias: Scientific foundations. *California Law Review*, *94*, 945–967.

Greenwald, A. G., McGhee, D. E., and Schwartz, J.L.K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*, 1464–1480.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., and Banaji, M. R. (2009). Understanding and using the implicit association test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*, 17–41.

Greenwald, A. G., Smith, C. T., Sriram, N., Bar-Anan, Y., and Nosek, B. A. (2009). Implicit race attitudes predicted vote in the 2008 U.S. presidential election. *Analyses of Social Issues and Public Policy*, *9*(1), 241–253. doi:10.1111/j.1530-2415.2009.01195.x

Gross, E, F., and Hardin, C. D. (2007). Implicit and explicit stereotyping of adolescents. *Social Justice Research*, *20*, 140–160.

Han, H. A., Olson, M. A., and Fazio, R. H. (2006). The influence of experimentally created extrapersonal associations on the Implicit Association Test. *Journal of Experimental Social Psychology*, *42*, 259–272.

Hardin, C. D., and Conley, T. D. (2001). A relational approach to cognition: Shared experience and relationship affirmation in social cognition. In G. B. Moskowitz (Ed.), *Cognitive social psychology: The Princeton Symposium on the Legacy and Future of Social Cognition* (pp. 3–17). Mahwah, NJ: Erlbaum.

Hardin, C. D., and Higgins, E. T. (1996). Shared reality: How social verification makes the subjective objective. In E. T. Higgins and R. M. Sorrentino (Eds.), *Handbook of motivation and cognition: The interpersonal context* (Vol. 3, pp. 28–84). New York: Guilford.

Hardin, C. D., and Rothman, A. J. (1997). Rendering accessible information relevant: The applicability of everyday life. In R. S. Wyer (Ed.), *The automaticity of everyday life: Advances in social cognition* (Vol. 10, pp. 143–156). Mahwah, NJ: Erlbaum.

Harrison, L. A., Stevens, C. M., Monty, A. N., and Coakley, C. A. (2006). The consequences of stereotype threat on the academic performance of white and non-white lower income college students. *Social Psychology of Education*, *9*, 341–357.

Henry, P. J., and Hardin, C. D. (2006). The contact hypothesis revisited: Status bias in the reduction of implicit prejudice in the United States and Lebanon. *Psychological Science*, *17*, 862–868.

Higgins, E. T. (1996). Knowledge activation: Accessibility, applicability, and salience. In E. T. Higgins and A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 133–168). New York: Guilford.

Higgins, E. T., Rholes, W. S., and Jones, C. R. (1977). Category accessibility and impression formation. *Journal of Experimental Social Psychology*, *13*, 141–154.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., and Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality and Social Psychology Bulletin*, *31*, 1369–1385.

Hugenberg, K., and Bodenhausen, G. V. (2003). Facing prejudice: Implicit prejudice and the perception of facial threat. *Psychological Science*, *14*, 640–643.

Jackman, M. (1994). *The velvet glove: Paternalism and conflict in gender, class, and race relations.* Berkeley, CA: University of California Press.

Jost, J. T., and Banaji, M. R. (1994). The role of stereotyping in system-justification and the production of false consciousness. *British Journal of Social Psychology*, *33*(1), 1–27. doi:10.1111/j.2044-8309.1994.tb01008.x

Jost, J. T., Banaji, M. R., and Nosek, B. A. (2004). A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political Psychology*, *25*, 881–919.

Jost, J. T., and Burgess, D. (2000). Attitudinal ambivalence and the conflict between group and system justification motives in low status groups. *Personality and Social Psychology Bulletin*, *26*, 293–305.

Jost, J. T., Pelham, B. W., and Carvallo, M. R. (2002). Non-conscious forms of system justification: Implicit and behavioral preferences for higher status groups. *Journal of Experimental Social Psychology*, *38*(6), 586–602. doi:10.1016/S0022-1031(02)00505-X

Jost, J. T., Rudman, L. A., Blair, I. V., Carney, D. R., Dasgupta, N., Glaser, J., and Hardin, C. D. (2009). The existence of implicit prejudice is beyond scientific doubt: A refutation of ideological and methodological objections and executive summary of ten studies no manager should ignore. *Research in Organizational Behavior*, *29*, 39–69.

Jost, J. T., and Thompson, E. P. (2000). Group-based dominance and opposition to equality as independent predictors of self-esteem, ethnocentrism, and social policy attitudes among African Americans and European Americans. *Journal of Experimental Social Psychology*, *36*, 209–232.

Kang, J. and Banaji, M. R. (2006). Fair measures: A behavioral realist revision of "affirmative action." *California Law Review*, *94*, 1063–1118.

Keifer, A. K., and Sekaquaptewa, D. (2007). Implicit stereotypes and women's math performance: How implicit gender-math stereotypes influence women's susceptibility

to stereotype threat. *Journal of Experimental Social Psychology*, *43*, 825–832.

Kim, D-Y. (2003). Voluntary controllability of the Implicit Association Test (IAT). *Social Psychology Quarterly*, *66*, 83–96.

Knowles, E. D., Lowery, B. S., and Schaumberg, R. L. (2010). Racial prejudice predicts opposition to Obama and his health care reform plan. *Journal of Experimental Social Psychology*, *46*(2), 420–423. doi:10.1016/j.jesp.2009.10.011

Kressin, N. R., and Petersen, L. A. (2001). Racial differences in the use of invasive cardiovascular procedures: Review of the literature and prescription for future research. *Annual Review of Internal Medicine*, *135*, 352–366.

Lambert, A. J., Payne, B. K., Ramsey, S., and Shaffer, L. M. (2005). On the predictive validity of implicit attitude measures: The moderating effect of perceived group variability. *Journal of Experimental Social Psychology*, *41*, 114–128.

La Pierre, R. (1934). Attitude vs action. *Social Forces*, *13*, 230–237.

Lee, S., Rogge, R. D., and Reis, H. T. (2010). Assessing the seeds of relationship decay. *Psychological Science*, *21*(6), 857–864. doi:10.1177/0956797610371342

Levy, B. (1996). Improving memory in old age through implicit self-stereotyping. *Journal of Personality and Social Psychology*, *71*, 1092–1107.

Lieberman, M. D. (2000). Intuition: A social cognitive neuroscience approach. *Psychological Bulletin*, *126*, 109–137.

Livingston, R. W. (2002). The role of perceived negativity in the moderation of African Americans' implicit and explicit racial attitudes. *Journal of Experimental Social Psychology*, *38*, 405–413.

Lowery, B. S., Hardin, C. D., and Sinclair, S. (2001). Social influence effects on automatic racial prejudice. *Journal of Personality and Social Psychology*, *81*, 842–855.

Maison, D., Greenwald, A. G., and Bruin, R. H. (2004). Predictive validity of the Implicit Association Test in studies of brands, consumer attitudes, and behavior. *Journal of Consumer Psychology*, *14*, 405–415.

Marsh, K. L., Johnson, B. L., and Scott-Sheldon, L. A. (2001). Heart versus reason in condom use: Implicit versus explicit attitudinal predictors of sexual behavior. *Zeitschrift fur Experimentelle Psychologie*, *48*, 161–175.

McConnell, A. R., and Liebold, J. M. (2002). Relations between the Implicit Association Test, explicit racial attitudes, and discriminatory behavior. *Journal of Experimental Social Psychology*, *37*, 435–442.

Meyer, D. E., and Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*, 227–234.

Mitchell, J. A., Nosek, B. A., and Banaji, M. R. (2003). Contextual variations in implicit evaluation. *Journal of Experimental Psychology: General*, *132*, 455–469.

Neely, J. H. (1976). Semantic priming and retrieval from lexical memory: Evidence for facilitatory and inhibitory processes. *Memory and Cognition*, *4*, 648–654.

———. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, *106*, 225–254.

Nock, M. K., and Banaji, M. R. (2007). Prediction of suicide ideation and attempts among adolescents using a brief performance-based test. *Journal of Consulting and Clinical Psychology*, *75*, 707–715.

Nock, M. K., Park, J. M., Finn, C. T., Deliberto, T. L., Dour, H. J., and Banaji, M. R. (2010). Measuring the suicidal mind. *Psychological Science*, *21*(4), 511–517. doi:10.1177/0956797610364762

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General*, *134*, 565–584.

Nosek, B. A., Banaji, M. R., and Greenwald, A. G. (2002a). Harvesting implicit group attitudes and beliefs from a demonstration website. *Group Dynamics*, *6*, 101–115.

———. (2002b). Math = male, me = female, therefore math ≠ me. *Journal of Personality and Social Psychology*, *83*, 44–59.

Nosek, M. A. (1995). Sexual abuse of women with physical disabilities. In T. N. Monga (Ed.), *Physical medicine and rehabilitation state of the art reviews: Sexuality and disability* (pp. 487–502). Philadelphia: Hanley Belfus.

Olson, M. A., and Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. *Psychological Science*, *12*, 413–417.

———. (2002). Implicit acquisition and manifestation of classically conditioned attitudes. *Social Cognition*, *20*, 89–104.

———. (2003). Relations between implicit measures of prejudice: What are we measuring? *Psychological Science*, *14*, 636–639.

———. (2004). Reducing the influence of extrapersonal associations on the Implicit Association Test: Personalizing the IAT. *Journal of Personality and Social Psychology*, *86*, 653–667.

———. (2006). Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin*, *32*, 421–433.

Orfield, G. (2001). *Schools more separate: Consequences of a decade of resegregation*. Cambridge, MA: Harvard University.

Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology*, *81*, 181–192.

Payne, B. K., Shimizu, Y., and Jacoby, L. L. (2005). Mental control and visual illusions: Toward explaining race-biased weapon identifications. *Journal of Experimental Social Psychology*, *41*, 36–47.

Penner, L. A., Dovidio, J. F., West, T. V., Gaertner, S. L., Albrecht, T. L., Dailey, R. K., and Markova, T. (2010). Aversive racism and medical interactions with Black patients: A field study. *Journal of Experimental Social Psychology*, *46*(2), 436–440. doi:10.1016/j.jesp.2009.11.004

Perdue, C. W., and Gurtman, M. B. (1990). Evidence for the automaticity of ageism. *Journal of Experimental Social Psychology*, *26*, 199–216.

Perugini, M. (2005). Predictive models of implicit and explicit attitudes. *British Journal of Social Psychology*, *44*, 29–45.

Perugini, M. O'Gorman, R., and Prestwich, A. (2007). An ontological test of the IAT: Self-activation can increase predictive validity. *Experimental Psychology*, *54*, 134–147.

Petersen, L. A., Wright, S. M., Peterson, E. D., and Daley, J. (2002). Impact of race on cardiac care and outcomes in veterans with acute myocardial infarction. *Medical Care*, *40*, 186–196.

Phelps, E. A., O'Conner, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., and Banaji, M. R. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience*, *12*, 1–10.

Plant, E. A., and Peruche, B. M. (2005). The consequences of race for police officers' responses to criminal suspects. *Psychological Science*, *16*, 180–183.

Plant, E. A., Peruche, B. M., and Butz, D. A. (2005). Eliminating automatic racial bias: Making race non-diagnostic for responses to criminal suspects. *Journal of Experimental Social Psychology*, *41*, 141–156.

Pletcher M. J., Kertesz, S. G., Kohn, M.A., and Gonzales, R. (2008). Trends in opioid prescribing by race/ethnicity for patients seeking care in US emergency departments. *Journal of the American Medical Association*, *299*, 70–78.

Putnam, R. D. (2007). *E Pluribus Unum*: Diversity and community in the twenty-first century, The 2006 Johan Skytte Prize Lecture. *Scandinavian Political Studies*, *30*, 137–174.

Ranganath, K. A., Smith, C. T., and Nosek, B. A. (2008). Distinguishing automatic and controlled components of attitudes from direct and indirect measurement. *Journal of Experimental Social Psychology*, *44*(2), 386–396.

Ratcliff, R., and McKoon, G. (1988). A retrieval theory of priming in memory. *Psychological Review*, *95*, 385–408.

———. (1994). Retrieving information from memory: Spreading activation theories versus compound cue theories. *Psychological Review*, *101*, 177–184.

Richeson, J. A., and Ambady, N. (2003). Effects of situational power on automatic racial prejudice. *Journal of Experimental Social Psychology*, *39*, 177–183.

Richeson, J. A., and Nussbaum, R. J. (2004). The impact of multiculturalism versus color-blindness on racial bias. *Journal of Experimental Social Psychology*, *40*(3), 417–423. doi:10.1016/j.jesp.2003.09.002

Richeson, J. A., and Shelton, J. N. (2005). Thin slices of racial bias. *Journal of Nonverbal Behavior*, *29*, 75–86.

Richeson, J. A., and Trawalter, S. (2005). Why do interracial interactions impair executive function? A resource depletion account. *Journal of Personality and Social Psychology*, *88*, 934–947.

Roccato, M., and Zogmaister, C. (2010). Predicting the vote through implicit and explicit attitudes: A field research. *Political Psychology*, *31*, 249–274.

Rooth, D-O. (2010). Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics*, *17*(3), 523–534.

Ross, L., and Nisbett, R. E. (1991). *The person and the situation: Perspectives of social psychology*. New York: McGraw-Hill.

Rudman, L. A., and Ashmore, R. D. (2007). Discrimination and the Implicit Association Test. *Group Processes and Intergroup Relations*, *10*, 359–372.

Rudman, L. A., Ashmore, R. D., and Gary, M. L. (2001). 'Unlearning' automatic biases: The malleability of implicit stereotypes and prejudice. *Journal of Personality and Social Psychology*, *81*, 856–868.

Rudman, L. A., and Borgida, E. (1995). The afterglow of construct accessibility: The behavioral consequences of priming men to view women as sexual objects. *Journal of Experimental Social Psychology*, *31*, 493–517.

Rudman, L. A., Feinberg, J. M., and Fairchild, K. (2002). Minority members' implicit attitudes: Ingroup bias as a function of ingroup status. *Social Cognition*, *20*, 294–320.

Rudman, L. A., and Glick, P. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues*, *57*, 743–762.

Rudman, L. A. and Goodwin, S. A. (2004). Gender differences in automatic ingroup bias: Why do women like women more than men like men? *Journal of Personality and Social Psychology*, *87*, 494–509.

Rudman, L. A., and Heppen, J. (2003). Implicit romantic fantasies and women's interest in personal power: A glass slipper effect? *Personality and Social Psychology Bulletin*, *29*, 1357–1370.

Rudman, L. A., and Lee, M. R. (2002). Implicit and explicit consequences of exposure to violent and misogynous rap music. *Group Processes and Intergroup Relations*, *5*, 133–150.

Rutland, A., Cameron, L., Bennett, L., and Ferrell, J. (2005). Interracial contact and racial constancy: A multi-site study of racial intergroup bias in 3–5 year old Anglo-British children. *Journal of Applied Developmental Psychology*, *26*(6), 699–713. doi:10.1016/j.appdev.2005.08.005

Sagar, H. A., and Schofield, J. W. (1980). Racial and behavioral cues in black and white children's perceptions of ambiguously aggressive acts. *Journal of Personality and Social Psychology*, *39*, 590–598.

Sears, D. O., and Henry, P. J. (2005). Over thirty years later: A contemporary look at symbolic racism. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 37, pp. 95–150). San Diego, CA: Academic Press.

Sekaquaptewa, D., Espinoza, P., Thompson, M., Vargas, P., and von Hippel, W. (2003). Stereotypic explanatory bias: Implicit stereotyping as a predictor of discrimination. *Journal of Experimental Social Psychology*, *39*(1), 75–82. doi:10.1016/S0022-1031(02)00512-7

Sherman, S. J., Presson, C. J., Chassin, L., Rose, J. S., and Koch, K. (2002). Implicit and explicit attitudes toward cigarette smoking: The effects of context and motivation. *Journal of Social and Clinical Psychology*, *22*, 13–39.

Shih, M., Ambady, N., Richeson, J. A., Fujita, K., and Gray, H. (2002). Stereotype performance boosts: The impact of self-relevance and the manner of stereotype activation. *Journal of Personality and Social Psychology*, *83*, 638–647.

Shih, M., Pittinsky, T. L., and Ambady, N. (1999). Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological Science*, *10*, 81–84.

Shook, N. J., and Fazio, R. H. (2007). *The effect of interracial versus same-race roommate relationships on attitudes.* Poster presented at the Annual Conference for the Society for Psychology and Social Psychology, Memphis, TN.

Sinclair, S., Hardin, C. D., and Lowery, B. S. (2006). Self-stereotyping in the context of multiple social identities. *Journal of Personality and Social Psychology*, *90*, 529–542.

Sinclair, S., Lowery, B. S., and Dunn, E. (2005). The relationship between parental racial attitudes and children's implicit prejudice. *Journal of Experimental Social Psychology*, *14*, 283–289.

Sinclair, S., Lowery, B. S., Hardin, C. D., and Colangelo, A. (2005). Social tuning of automatic racial attitudes: The role of affiliative motivation. *Journal of Personality and Social Psychology*, *89*, 583–592.

Skinner, B. F. (1953). *Science and human behavior.* New York: Macmillan.

Sniderman, P. M., and Carmines, E. G. (1997). *Reaching beyond race.* Cambridge, MA: Harvard University Press.

Spalding, L. R., and Hardin, C. D. (1999). Unconscious unease and self-handicapping: Behavioral consequences of individual differences in implicit and explicit self-esteem. *Psychological Science*, *10*, 535–539.

Spencer, S. J., Steele, C. M., and Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*, 4–28.

Srull, T. K., and Wyer, R. S. (1979). The role of category accessibility in the interpretation of information about persons: Some determinants and implications. *Journal of Personality and Social Psychology*, *38*, 841–856.

Steele, C. M., and Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*, 797–811.

Steffens, M. C. (2004). Is the Implicit Association Test immune to faking? *Experimental Psychology*, *51*, 165–179.

Swanson, J. E., Rudman, L. A., and Greenwald, A. G. (2001). Using the Implicit Association Test to investigate attitude-behavior consistency for stigmatized behavior. *Cognition and Emotion*, *15*, 207–230.

Sylvestre, D. L., Litwin, A. H., Clements, B. J., and Gourevitch, M. N. (2005). The impact of barriers to hepatitis C virus treatment in recovering heroin users maintained on methadone. *Journal of Substance Abuse Treatment*, *29*, 159–165.

Tam, T., Hewstone, M., Cairns, E., Tausch, N., Maio, G., and Kenworthy, J. B. (2007). The impact of intergroup emotions on forgiveness in Northern Ireland. *Group Processes and Intergroup Relations*, *10*, 119–135.

Tam, T., Hewstone, M., Harwood, J., Voci, A., and Kenworthy, J. (2006). Intergroup contact and grandparent-grandchild communication: The effects of self-disclosure on implicit and explicit biases against older people. *Group Processes and Intergroup Relations*, *9*, 413–430.

Teachman, B. A., Marker, C. D., and Smith-Janik, S. B. (2008). Automatic associations and panic disorder: Trajectories of change over the course of treatment. *Journal of Consulting and Clinical Psychology*, *76*(6), 988–1002.

Teachman, B. A., Smith-Janik, S. B., and Saporito, J. (2007). Information processing biases and panic disorder: Relationships among cognitive and symptom measures. *Behaviour Research and Therapy*, *45*, 1791–1811.

Teachman, B. A., and Woody, S. R. (2003). Automatic processing in spider phobia: Implicit fear associations over the course of treatment. *Journal of Abnormal Psychology*, *112*, 100–109.

Tetlock, P. E., and Mitchell, G. (in press). Unconscious prejudice and accountability systems: What must organizations do to check implicit bias? *Research in Organizational Behavior.*

Turner, R. N., Hewstone, M., and Voci, A. (2007). Reducing explicit and implicit outgroup prejudice via direct and extended contact: The mediating role of self-disclosure and intergroup anxiety. *Journal of Personality and Social Psychology*, *93*(3), 369–388. doi:10.1037/0022-3514.93.3.369

U. S. Department of Justice. (2001). *Policing and homicide, 1976–98: Justifiable homicide by police, police officers murdered by felons* (NCJ 180987). Washington, DC: Bureau of Justice Statistics.

van Breukelen, W., van der List, R., and Steensma, H. (2004). Voluntary employee turnover: Combining variables from the "traditional" turnover literature

with the theory of planned behavior. *Journal of Organizational Behavior*, *25*, 893–914.

von Hippel, W., Brener, L., and von Hippel, C. (2008). Implicit prejudice toward injecting drug users predicts intentions to change jobs among drug and alcohol nurses. *Psychological Science*, *19*, 7–11.

Wegener, D. T., and Petty, R. E. (1997). The flexible correction model: The role of naive theories of bias in bias correction. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 29, pp. 141–208). San Diego, CA: Academic Press.

Weirs, R. W., Woerden, N. V., Smulders, F. T., and de Jong P. T. (2002). Implicit and explicit alcohol-related cognitions in heavy and light drinkers. *Journal of Abnormal Psychology*, *111*, 648–658.

Wittenbrink, B., Judd, C. M., and Park, B. (1997). Evidence for racial prejudice at the implicit level and its relationship to questionnaire measures. *Journal of Personality and Social Psychology*, *72*, 262–274.

Wittenbrink, B., Judd, C. M., and Park, B. (2001). Spontaneous prejudice in context: Variability in automatically activated attitudes. *Journal of Personality and Social Psychology*, *81*(5), 815–827. doi:10.1037/0022-3514.81.5.815

Wittenbrink, B., and Schwarz, N. (Eds.) (2007). *Implicit measures of attitudes*. New York: Guilford.

Word, C. O., Zanna, M. P., and Cooper, J. (1974). The nonverbal mediation of self-fulfilling prophecies in interracial interaction. *Journal of Experimental Social Psychology*, *10*, 109–120.

Wraga, M., Helt, M., Jacobs, E., and Sullivan, K. (2007). Neural basis of stereotype-induced shifts in women's mental rotation performance. *Social Cognitive and Affective Neuroscience*, *2*, 12–19.

Ziegert, J. C., and Hanges, P. J. (2005). Employment discrimination: The role of implicit attitudes, motivation, and a climate for racial bias. *Journal of Applied Psychology*, *90*, 553 –562.

CHAPTER 2

Biases in Interracial Interactions

Implications for Social Policy

J. NICOLE SHELTON

JENNIFER A. RICHESON

JOHN F. DOVIDIO

The *Brown v. Board of Education* decision in 1954 and the Civil Rights Act of 1964 were monumental policy decisions that changed the landscape of race relations in the United States. Before the implementation of these policies, ethnic minorities and Whites had very little contact with one another, primarily because ethnic minorities were not allowed to be in the same settings with Whites, including attending the same schools, working in the same place of employment, living in the same neighborhoods, and even riding in the same sections of buses and eating in the same sections of restaurants. When contact between the groups occurred, it was often fraught with extreme hostility, fear, and anxiety. Without a doubt, the policy decisions of 1954 and 1964 against segregation and discrimination increased the opportunity for people to engage in contact with members outside of their racial group. In addition, these policy decisions paved the way for improvements in social norms toward ethnic minorities, such that it was unacceptable to publicly express negative racial beliefs or behave in a discriminatory manner toward them. The change in social norms eventually improved individuals' private attitudes and behaviors. Indeed, there has been a substantial increase in Whites' endorsement of racial equality and integration over the past fifty-plus years (Bobo, 2001).

Although federal laws and organizational policies have been developed and have been relatively effective in reducing blatant forms of bias (Fiske and Krieger; Hardin and Banaji, this volume), it is more challenging to create laws and policies to reduce the subtle bias that is often present in everyday interracial interactions. For example, a law cannot be created that prohibits Whites from displaying negative nonverbal behaviors toward African Americans during daily interactions. Yet, subtle negative behavior and signals can have adverse effects on the performance and ambitions of African Americans (Purdie-Vaughns et al., 2008; Salvatore and Shelton, 2007). Moreover, efforts to demand that individuals comply with egalitarian social norms can provoke hostility and have the unintended consequence of worsening the problems (Plant and Devine, 2001). Given the problems that both blatant and subtle bias create, it is important to understand how both types influence everyday interracial interactions and how policies may help improve the quality of interracial interactions and ultimately reduce racial bias.

In this chapter, we explore how and why racial bias systematically influences daily interracial interactions across three contexts: (a) residential spaces on college campuses, (b) health care provider-patient dyads in medical settings, and (c) employee-employer relations in the workplace. We focus our attention specifically on these contexts, in part, because the landmark policy decisions of 1954 and 1964 opened the doors for increased opportunities for individuals to engage in contact across the racial divide in these three areas. Moreover, we selected these contexts because they are venues into which most people will enter at some point during their lives. With respect to residential experiences in colleges, we are aware, of course, that not everyone attends college. However, given that policies about the American educational system played such a profound role in creating spaces for intergroup contact, it seems essential to explore the dynamics of contact in this setting. In addition, some of the most influential work on contact theory occurred in residential areas and revealed that contact was related to improved intergroup attitudes (e.g., Deutsch and Collins, 1951; Wilner, Walkley, and Cook, 1955). Moreover, studying roommate relationships on college campuses may offer insight into ways to improve the academic outcomes of students (see Cohen and Garcia, this volume for policies about the racial achievement gap).

Regardless of educational level, the majority of people in the United States will interact with some form of medical-care provider and will obtain employment at some point in their lives. Racial bias in these contexts has the potential to inhibit successful life outcomes. Health-care providers' racial biases, for example, have been posited to contribute to racial disparities in health outcomes (Dovidio et al., 2008). The Institute of Medicine's 2003 report on unequal medical treatment acknowledge that such biases, if they exist, undermine the high ethical standards that medical professionals are accountable for upholding

(Smedley, Stith, and Nelson, 2003).[1] Describing how racial biases interfere with the behaviors of a group of well-meaning people who have taken an oath to treat all individuals equally highlights the insidious nature of racial bias. Understanding when and how racial biases may influence racial disparities in health care outcomes may begin to shed light on policies that could be developed to improve the life expectancy of ethnic minorities.

Finally, we focus on the workplace because, although there are explicit federal laws and sanctions against blatant racial bias in the workplace, many organizational-level practices are such that subtle bias remains a problematic force. These practices trickle down to influence the ways in which Whites and ethnic minorities interact with one another in their daily lives while at work. Taken together, these three contexts are common avenues for interracial contact to occur and are ones in which policies and regulations have been put forth, or could be established, to improve the quality of Whites' and ethnic minorities' life experiences.

The chapter is divided into three primary sections. In the first, we provide an overview of the literature on contemporary intergroup bias, particularly with respect to race in the United States. In the second, we discuss the processes associated with the interplay between racial attitudes and interracial contact. And in the third, we discuss the implications of these processes for the aforementioned three contexts, paying special attention to how policies in these contexts may shape individuals' experiences during interracial encounters and how knowing individuals' experiences may offer ideas about policy decisions. We pay particular attention to the fact that racial biases may have different consequences, in some cases completely opposite effects, for Whites and ethnic minorities during their interactions in these settings. These divergent experiences pose a major challenge for policy: policies must be tailored in such a way that an improvement in the lives of one group does not cause harm in the lives of the other. In essence, we explore how policy decisions shape the psychology of interracial interactions and how the psychology of interracial interactions may shape policies established to create harmonious interracial relations. Because we cover a lot of ground by describing interracial contact across three settings, our review of the literature is not meant to be exhaustive. Instead, we sample a few classic and contemporary articles that highlight the issues of concern in the best light. In addition, we focus primarily on race relations in the United States within the context of Black-White relations, which have historically been central to the development of social policy defining intergroup relations in the United States. Finally, we offer a synthesis of common practices based on intergroup contact research that policy makers use to maximize the benefits of diversity across multiple settings.

Intergroup Bias

Intergroup bias, a pervasive and arguably universal phenomenon within and across many cultures (Sidanius and Pratto, 1999), stems from processes associated with prejudice and stereotyping. Prejudice reflects a general negative evaluation of a group, whereas stereotyping reflects the association of specific traits to a group. Prejudice and stereotypes often lead to discrimination, which is the unjustified group-based difference in behavior that gives one group an advantage over others. Perhaps intergroup bias is a pervasive phenomenon because there are several normal processes that allow people to navigate a complex environment that predispose them to developing intergroup prejudices. For example, the ability to sort people, spontaneously and with minimum effort and awareness, into meaningful categories is a universal facet of human perception essential for efficient functioning (Bodenhausen, Todd, and Becker, 2007). Given the importance of the self in social perception, social categorization further involves a basic distinction between the group containing the self (in-group) and other groups (out-groups)—or between the "we's" and the "they's" (Turner et al., 1987). The recognition of different group memberships shapes social perception, affect, cognition, and behavior in ways that systematically produce intergroup biases (Gaertner and Dovidio, 2000). If, when, and how bias is manifested, however, depends upon cultural norms, individual motivation, the historical relations between groups, and the immediate circumstances (Crandall and Eshleman, 2003). In societies that place high value on egalitarianism, going as far as establishing laws to promote equality, intergroup biases often take the form of subtle, rather than blatant, prejudice (Gaertner and Dovidio, 1986).

The discrepancy between the ideal of egalitarianism and the psychological forces that promote racial bias has been posited as a critical factor leading to the development of subtle forms of racial bias. Whereas the traditional form of racial bias represented the overt expression of dislike and hostility, as well as the endorsement of negative cultural stereotypes, contemporary forms of racial bias involve more complex dynamics and typically more subtle expressions of bias. This is evident in research within a framework of *aversive racism* (Dovidio and Gaertner, 2004), which

shows that most Whites who express egalitarian values and nonprejudiced attitudes also, because of basic principles associated with categorization that promote bias, harbor unconscious negative feelings and beliefs about African Americans. These unconscious feelings and beliefs develop through both common cultural experiences (e.g., media exposure that promotes stereotypes) and fundamental psychological processes (e.g., favoritism for members of one's own group). As a consequence, implicit measures of attitudes (e.g., the Implicit Association Test; see Hardin and Banaji, this volume) reveal that, even though most White Americans indicate that they are not prejudiced on self-report measures, the majority of White Americans possess these unconscious racial biases.

The distinction and general dissociation of explicit and implicit attitudes, which characterizes aversive racism, has important implications for the way the discrimination is manifested. Explicit and implicit attitudes influence behavior in different ways (Fazio, 1990; Hardin and Banaji, this volume). Explicit attitudes shape deliberative, well-considered responses for which people have the motivation and opportunity to weigh the advantages and disadvantages of their responses. Implicit attitudes influence responses that are more difficult to monitor and control or responses that people do not view as an indication of their attitude and, thus, do not try to control. For example, Whites' explicit racial attitudes tend to predict their trait ratings of African Americans, whereas their implicit racial attitudes tend to predict how much they smile at African Americans during an interaction. In general, because aversive racists do not perceive themselves as racists but possess implicit negative racial attitudes, they are not likely to engage in blatant forms of racial bias, such as treating African Americans in a negative manner when it is clear that only race could be used as an explanation. However, aversive racists do engage in more subtle forms of bias, such as avoiding African Americans or displaying negative nonverbal behaviors, especially in less structured situations where the norms for the appropriate behavior are more ambiguous or there are non-race-related factors that could be used to explain their behavior.

Although there is a distinction between the types of behaviors the different types of attitudes predict—explicit predict verbal behaviors and implicit predict nonverbal behaviors—the extent to which both types of racial attitudes predict behaviors during interactions depends on peoples' motivation to control their responses. Both types of attitudes tend to predict behaviors best when people are relatively unmotivated, or lack the cognitive resources, to control their responses. In these cases, the more negative individuals' racial attitudes, the more negative their behaviors and judgments. For example, when White Americans are low in motivation to control prejudiced reactions, the more negative are their implicit racial attitudes, the more negative are their ratings of a typical African American male student (Dunton and Fazio, 1997), the more negative are traits they assign to African American targets relative to White targets during a first impression task (Olson and Fazio, 2004), and the more discomfort they anticipate experiencing with African Americans (Towles-Schwen and Fazio, 2003). When White Americans are motivated to control their reactions, however, they make more positive judgments of African Americans, even if their implicit attitudes are negative.

Intergroup Contact

This analysis of the psychological complexity of intergroup bias offers valuable insights into understanding the dynamics of interracial contact. Historically, appropriately structured intergroup contact has represented psychology's main remedy for reducing prejudice. Allport (1954) proposed that mere, or superficial, contact with out-group members would not necessarily reduce intergroup bias but instead may reinforce stereotypes and initial suspicion. He argued that contact with out-group members improves intergroup attitudes under the right conditions: specifically, when (a) there is equal status between the group members in the particular contact situation, (b) group members have common goals, (c) there is a high level of interdependence and cooperation among group members, and (d) contact is encouraged and supported by authorities, customs, and laws.

Research on intergroup contact played a large role in the 1954 Supreme Court school desegregation decision in *Brown v. Board of Education*. In the psychological briefs referring to relations between Whites and African Americans, researchers argued that, "Segregation leads to a blockage in the communication and interaction between the two groups. Such blockages tend to increase mutual suspicion, distrust, and hostility" (*Brown v. Board of Education*, as cited in Martin, 1998, p. 145). The ruling made in the *Brown v. Board of Education* court case would pave the way for Allport's (1954) conditions to be implemented in the United States educational system. Since then, an extensive body of research has been conducted on how intergroup contact is an antidote for reducing intergroup bias, and a meta-analysis of 515 studies revealed that intergroup contact is associated with lower levels of intergroup bias across many types of target groups (e.g., racial and ethnic groups, heterosexuals and gays and lesbians, elderly and young

adults, disabled and nondisabled) (Pettigrew and Tropp, 2006).

Despite the generally impressive support for contact theory that has accumulated over the years, recent work has also identified an important qualifying factor: majority and minority group members respond to intergroup contact in different ways. Specifically, the relationship between contact and more favorable intergroup attitudes is weaker for ethnic minorities than for Whites (Tropp and Pettigrew, 2005). In fact, some research has shown that African Americans who have had greater contact with Whites tend to have more negative attitudes toward Whites, largely due to their perceptions of Whites' level of bias toward African Americans (Livingston, 2002). Given this difference, it is essential for policy makers to take into consideration that solutions that work well for one group may be less effective for others. Thus, understanding the causes of the different reactions of Whites and racial minorities can critically inform the development of social policies.

Much of the work on contact theory has focused on the conditions under which contact occurs (e.g., equal status for participants). We suggest that, in addition, it is important to understand how the complexity of contemporary intergroup bias can influence the affective, cognitive, and behavioral outcomes during interactions. The distinction between explicit and implicit racial attitude, as noted previously, shows how Whites and minorities may have divergent experiences during interracial interactions. For example, Dovidio, Kawakami, and Gaertner (2002) found that Whites who explicitly reported that they had more negative racial attitudes behaved in a less verbally friendly way toward an African American compared to a White partner. However, it was Whites' implicit racial attitudes that predicted how biased their less controllable, nonverbal behaviors were. In other words, Whites tended to send mixed messages to their African American partners that were composed of positive verbal, but negative nonverbal, behaviors. Interestingly, however, African Americans formed their impressions of their White partners' friendliness from those partners' nonverbal behaviors, which were largely negative, causing African Americans to have an unfavorable impression of their White partner.

Furthermore, the mixed messages Whites express during interracial interactions can have detrimental consequences for minorities' experiences during the interaction. Salvatore and Shelton (2007), for example, found that African Americans were more cognitively depleted after exposure to subtle, compared to blatant, racial bias in a job hiring setting. Similarly, Dovidio (2001) found that it took dyads involving an African American and White person categorized as an aversive racist longer to solve a problem than dyads involving an African American and a blatantly racist White. Presumably, the mixed messages and subtle racial bias displayed by the aversive racist interfered more with the effectiveness of solving the problem during the interaction than the straightforward negative behavior displayed by the blatantly racist White person.

In the remaining sections of this chapter, we consider ways in which findings about racial biases might guide policies in improving the ways in which individuals interact with one another across racial lines in three different contexts. We focus primarily on policy recommendations that stem from research on making people aware of their racial biases (Monteith and Mark, 2005) and on the Common Ingroup Identity Model (Gaertner and Dovidio, 2000).

Racial Bias in Context

In previous sections, we described the complex nature of contemporary racial attitudes and illustrated how it can produce divergent impressions and reactions in interracial interactions. Whereas those earlier sections illustrate the potential problems that can arise in relatively brief, typically socially oriented situations, in this section of the chapter we examine the implications of these processes for longer-term contact in settings that have profound impact on people's lives.

Residential Life in an Educational Context

With legal barriers to educational opportunities removed, Whites and ethnic minorities are now allowed to attend the same colleges and universities. In fact, the implementation of affirmative action programs as a means of redressing racial discrimination contributed to an increase in ethnic minorities at predominately White colleges, opening the doors for contact to occur between different racial groups in an educational context. Nevertheless, in many ways, universities remain functionally segregated, especially with respect to social settings. White and ethnic minority students frequently avoid contact with one another, sometimes because of racial bias and intergroup anxiety (Plant and Devine, 2003) and other times because they think the out-group does not want to have contact with them (Shelton and Richeson, 2005). One area, however, in which actual interactions occur, not just mere exposure to out-group members, is residential space on college campuses.

Residential housing on college campuses provides a rich living laboratory in which to study the experience of intergroup contact and how policies play an

important role in shaping daily race relations. The primary conditions posited by Allport (1954) as the ideal factors of intergroup contact to reduce prejudice are generally met in college residential housing situations. Specifically, students are of equal status—they are peers in the college environment. They generally have the common goal of making their living arrangement pleasant and comfortable, which means they are likely to be willing to work together for the greater good of the dyad. Moreover, they are interdependent, meaning that their behaviors have repercussions for one another. Finally, administrators and policy makers at universities tend to encourage interactions across racial and ethnic lines, and most universities have formal policies against using race as a factor in roommate assignments to avoid racial segregation. But do these policies at universities about residential spaces facilitate or hinder interracial harmony among students? Indeed, despite universities' attempts to diversify living arrangements, students often opt to self-segregate in dormitories and other social spaces on campus (Sidanius et al., 2004). What, then, are the consequences of these university policies, socially and personally for students, a substantial portion of whom find themselves in their most intimate contact with out-group members?

Psychological research suggests that there may be short-term costs but long-term benefits of diverse living arrangements on college campuses, and the costs and benefits may depend on the race of the residents. Consistent with basic research showing interracial interactions are more stressful and cognitively effortful than same-race interactions (see Hebl and Dovidio, 2005; Richeson and Shelton, 2007 for reviews), interracial roommate pairs are at greater risk for strained and discordant interactions than same-race roommate pairs. Research reveals, for example, that among randomly assigned roommates, Black-White roommate pairs are less satisfied with their living arrangement than are same-race roommate pairs (Phelps et al., 1996, 1998). Similarly, White freshmen who have been randomly assigned to have an African American roommate spend less time with their roommate, perform fewer activities together, and have less overlapping social-network involvement than their counterparts with a White roommate (Shook and Fazio, 2008; Towles-Schwen and Fazio, 2006). Moreover, both Whites and ethnic minorities (African Americans and Latinos) experience more anxiety, feel less authentic (i.e., believe they can not be themselves), like their roommate less, are more likely to wish they had a different roommate, and are less likely to want to live with their roommate again next year when they have been randomly assigned to a roommate not of their own racial group than when they have a same-race roommate (Shelton, Richeson, and Salvatore, 2005; Trail, Shelton, and West, 2009; West, Shelton, and Trail, 2009).

Furthermore, the negative consequences of having a White roommate can become worse over time for ethnic minorities. Trail, Shelton, and West (2009) used a daily diary procedure in which they followed cross-race and same-race roommate pairs across fifteen days near the beginning of the academic year. As a result, they were able to assess if any of the roommates' daily experiences changed across time for either group. Indeed, they did, but only for ethnic minorities. In the cross-race roommate pairs, the ethnic minority students' daily experiences worsened across the fifteen days. For example, ethnic minorities with White roommates experienced less positive moods than their White roommates, and their positive mood diminished over the course of the study. Similarly, ethnic minorities with White roommates perceived their roommates as behaving in a less positive manner toward them compared to the other roommate pairs, and these perceptions worsened over time. Interestingly, these outcomes improved over time for ethnic minorities with ethnic minority roommates. For example, ethnic minorities' positive mood increased over time when they had ethnic minority roommates. Taken together, these findings reveal that students in cross-race roommate arrangements often have more negative experiences than those in same-race roommate arrangements. More disturbingly, some of these negative experiences become worse over time for ethnic minorities. These findings are quite disheartening given that repeated contact under the right conditions—which, as stated previously, the college roommate situation has—should improve race relations.

Research also shows that explicit and implicit racial bias play pivotal roles in the dynamics of interracial roommate relationships. Shelton and Richeson (2005), for example, asked ethnic minority students (Blacks, Latinos, and Asians) who had been randomly assigned to have a White or another ethnic minority roommate to complete a daily diary measure about the quantity and quality of the roommate interactions over the course of three weeks. They found that among ethnic minorities with White roommates, the more negative their explicit racial attitudes toward Whites were, the more the minorities tended to avoid contact with their roommates. Moreover, as time went on, these ethnic minority students wanted less and less contact with their White roommates. Additionally, among ethnic minorities with White roommates the more negative their racial attitudes, the less close they felt to their roommate. Moreover, this lack of connection grew worse across the three weeks of the study. Also, among ethnic minorities with White

roommates, the more negative their racial attitudes, the less positive affect and the more negative affect they experienced with respect to their interactions with their roommate. Among ethnic minorities with ethnic minority roommates, racial attitudes were not related to their experiences with their roommate. Similarly, Towles-Schwen and Fazio (2006) showed that implicit racial attitudes are related to cross-race roommate experiences. Specifically, they examined the extent to which Whites' implicit racial attitudes predicted the longevity of their roommate relationship and how satisfied Whites were with the relationship. Whites' implicit racial attitudes predicted the longevity of the relationship, such that the more negative their attitudes, the more likely their relationship with a Black roommate dissolved by the end of the year. Thus, the racial attitudes of both ethnic minorities and Whites, as assessed by explicit and implicit measures, influence the quality of interracial roommate relationships.

If policy makers made decisions based on the daily experiences that students have as a result of living with an out-group member, they might decide that segregation should be the law. However, additional research shows that negative experiences are not a necessary outcome for roommates of different races or ethnicities. Although social categorization, which generally occurs along racial group lines in the United States, can form a basis for intergroup biases that can impair relationships between roommates, people can rely on other forms of social categorization. Gaertner and Dovidio (2000), for instance, have shown that when people feel that they share a common in-group identity (e.g., as students at the same college) with people otherwise seen only in terms of different (e.g., racial or ethnic) group memberships, they have more positive orientations toward them. Consistent with this position, West et al. (2009) found that when students had a strong belief that members of different racial and ethnic groups on campus shared a common university identity, roommates from different groups maintained feelings of friendship across their first month together, and at the end of the month had stronger friendships than did roommates of the same race or ethnicity. These effects occurred equivalently for White and racial/ethnic minority students. Thus, cross-group roommate relations can, under some conditions, be more positive than same-group roommate relations.

In determining policy, in addition to the direct impact on the roommates themselves, the broader, long-term benefits of diversified living arrangements on college campuses also need to be considered. Psychological research provides compelling evidence that mixed-race roommate relationships improve intergroup relations over time. Shook and Fazio (2008) administered White freshmen an implicit measure of racial bias and an Intergroup Anxiety Toward Blacks Scale during the first two weeks of an academic quarter and again during the last two weeks of the quarter. Results revealed that implicit racial attitudes became more positive and intergroup anxiety decreased over time for Whites who had been randomly assigned to have a Black compared to those assigned a White roommate. Furthermore, in one of the most impressive longitudinal studies on this issue to date, Van Laar et al. (2005) examined the causal relationship between university roommate arrangements and racial attitudes among four groups—Asians, Blacks, Latinos, and Whites. Using a five-wave, four-year panel study, they examined changes in racial attitudes among students who were randomly assigned during their first year of college and who volunteered in later years to live with out-group and in-group roommates. Overwhelmingly, the data revealed that being randomly assigned to have out-group roommates caused improvements in racial attitudes for all individuals. Taken together, this is strong evidence that living with out-group members during the college years is associated with improvements in racial attitudes for both Whites and ethnic minorities.

One of the advantages of Van Laar et al.'s (2005) dataset is that it allowed the researchers to examine the impact of both respondent's and roommate's ethnicity on changes in racial attitudes. Thus, not only could the researchers examine the general effects of having out-group roommates versus in-group roommates, as noted above, but they could also examine whether the specific race of the students had different effects on racial attitudes. The findings revealed that it did, primarily for respondents with an Asian roommate. That is, contrary to the intergroup contact hypothesis, respondents who lived with an Asian American roommate had more negative racial attitudes at a later time, especially if the respondent was White. Specifically, living with an Asian roommate was associated with decreased positive affect toward Blacks and Latinos, as well as increased intergroup anxiety and symbolic racism. This pattern was true regardless of whether students were randomly assigned to live with Asians during their first year of college or volunteered to live with Asians during their second and third year of college. The researchers suggested that this unfortunate pattern of results might have occurred because of the combination of peer socialization and Asians' more highly prejudiced attitudes. That is, the Asian students in their sample had significantly higher prejudiced beliefs on various measures of prejudice than the other three ethnic groups. Given that peer socialization studies show that people change their attitudes

and behavior to be consistent with their peers, it is highly likely that individuals shifted their attitudes to fit in with the beliefs of their prejudiced Asian roommate. These data show the importance of understanding that the ethnicity of the person one comes into contact with is essential for designing policies about residential life.

Given the sometimes short-term costs (i.e., negative daily interactions) but long-term benefits (i.e., improved racial attitudes) of racially diverse living arrangements among college students, what institutional policies should be set forth? One solution colleges have developed to deal in part with these costs and benefits has been to create theme floors and/or dorms that allow ethnic minorities (as well as other groups) to live together. It would appear, however, that this solution would solve the short-term costs (e.g., more pleasant roommate experiences) but inhibit the long-term benefits (e.g., racial tolerance). To our knowledge there has not been a systematic examination of the consequences of these arrangements. However, research on this issue might consider effects over time. For example, minorities on predominantly White campuses may prefer, and benefit in some ways by, living in theme housing with a substantial portion of minority residents. After making the personal transition to college, though, they may seek other forms of campus housing that enhances the amount and type of intergroup contact that they experience.

From a policy perspective, one question that has been raised as a result of these themed living arrangements is whether or not residential areas for certain racial groups are against most universities' antidiscriminatory policies. Perhaps believing that these residences are discriminatory, in 2006 the University of Massachusetts-Amherst began phasing out such residential areas, along with other race-conscious programs for minorities (Associated Press, 2006). The vice chancellor of student affairs and campus life at the University of Massachusetts–Amherst stated, "Students who come to the university need to be exposed to different opinions and ideas. When you have segregated pockets in our residence halls, we are allowing students to shut themselves off, and then they are missing out" (Associated Press, 2006). The University of Massachusetts–Amherst, as well as other universities across the nation, changed their policies so that the theme dorms/floors could not be exclusively for ethnic minorities but instead had to be for all individuals who are interested in learning about a particular culture. For example, White students are allowed to live in the Black theme dorm, and Black students are allowed to live in the Asian theme dorm. Given that students who select to live in these "program" dorms are most probably open to diversity, it is likely that the student interactions are more pleasant and conducive to learning regardless of race. Therefore, it is not clear if improvements in intergroup bias would occur.

Another policy-related question is that, given the major potential tension between the short-term problems and the long-term benefits associated with diverse college residential living arrangements, how can university officials design and implement policies that address both the levels of comfort students need to feel on a daily basis in their dorms and the educational and democratic benefits of diversified living arrangements? We suggest that any such policy recognize that there are unique challenges in cross-group roommate relationships beyond those that occur for roommates of the same race/ethnicity. We identify two such approaches rooted in the fundamental importance of social categorization in social relations. As we explained earlier, merely categorizing people into racial groups can breed negative feelings toward the out-group and foster in-group favoritism. Thus, policies and interventions might focus on changing the ways roommates from different racial/ethnic groups categorize each other. Specifically, one focus of policies and interventions might be to encourage *decategorization*, that is, reducing reliance on racial group membership in social perception by emphasizing the unique qualities of different people and promoting personalized interactions through self-disclosure. Policy initiatives, for example, can be designed to create opportunities for roommates to get to know one another and become friends *prior* to living together.

Research has shown that reciprocal personal self-disclosure and working together on shared leisure activities is a way to increase friendship and intimacy (Reis and Shaver, 1988). Building upon this idea, Aron et al. (1997) developed a "fast friend" paradigm in which pairs of individuals answer a series of questions that becoming increasingly more personal and also engage in relationship building tasks (e.g., play a game) together. Remarkably, pairs who engage in this fast friend task feel closer and more connected to one another than pairs who simply engage in small talk. Recently, this task has been used to reduce racial prejudice and create closeness among out-group members. For example, this paradigm was successful in building trust and admiration between police officers and Black community members (Aron et al., 2007). Moreover, Page-Gould, Mendoza-Denton, and Tropp (2008) had mixed-race and same-race strangers engage in the fast friend task, but adapted it to occur across three days. After participating in the task, participants completed a daily diary for ten consecutive days to assess the number of interracial interactions they initiated during that time and the

amount of conflict they experienced during those interactions. The results revealed that individuals' feelings of how close they felt to their out-group partner increased significantly across the fast friend sessions and ultimately reduced the stress individuals experienced, as measured by self-report and physiological (i.e., cortisol) measures. Moreover, the fast friend manipulation influenced the quantity and quality of individuals' interracial interactions in general, particularly for the individuals for whom these interactions were the most stressful. Specifically, highly prejudiced Whites who had made a cross-group friend initiated more interracial interactions during the follow-up ten days. Moreover, ethnic minorities who tended to believe that Whites would reject them on their basis of race had fewer interracial interactions that involved conflict during the post-ten-day diary period when they had made a cross-group, compared with a same-race, friend through the fast friend paradigm. Based on these studies, we recommend that the fast friend procedure be implemented during first-year orientation among all roommates, but especially among mixed-race roommate pairs.

Another approach would be to foster recategorization, replacing the focus on separate racial group identities with a salient common group identity. According to the Common Ingroup Identity Model (Gaertner and Dovidio, 2000), when members of different groups recategorize themselves into a single superordinate group instead of perceiving one another as "we's" versus "them's" attitudes and behaviors toward the former out-group members become more positive. In this case, if White and ethnic minorities focus on their common group membership as students of their universities (e.g., Princeton students) instead of as Whites and Blacks, then the dynamics of their daily interactions are likely to become more positive. As we reported earlier, West et al. (2009) found that when roommates from different racial/ethnic groups had a strong perception of common university identity across group lines, they established and maintained high levels of friendship over their first month on campus. Common group identity can be achieved through activities that repeatedly emphasize existing shared memberships (e.g., the same university or residence) or through cooperative activities (e.g., having roommates cooperate to achieve a number of goals during first-year orientation). Moreover, once a common identity is established, roommates are likely to engage in the reciprocal behaviors (e.g., mutual helping and disclosure) that can create a behavioral foundation for a positive relationship and ultimately produce more personalized interactions over time. Thus, recategorization and decategorization can operate in complementary ways as roommate relationships develop. In addition, because friendship with a member of another group is one of the most potent forms of intergroup contact, these positive roommate relationships can have cascading effects by improving intergroup attitudes more generally (Pettigrew, 1997).

Our basic premise in this chapter is that the complex and often subtle nature of contemporary intergroup bias can have widespread impact on intergroup interactions, and ultimately, on the outcomes for members of different racial and ethnic groups. The work on intergroup roommate relationships illustrates the fragility of intergroup relationships in situations of sustained and socially intimate intergroup contact. In the next section we examine how these same intergroup biases can also exert an adverse influence in task-oriented contact: medical encounters.

Medical Settings

At some point in life, people are likely to become ill and need to visit a medical facility. Prior to the 1964 civil rights legislation and Medicare and Medicaid legislation in 1965, ethnic minorities tended not to visit the same medical facilities as Whites. Today, long after such restrictions have been removed, ethnic minorities receive health care in predominately White facilities and interact with White physicians. However, they do not always receive the same high-quality medical treatment as Whites, nor do they have the same successful interactions with physicians as do Whites. In one controversial study (Schulman et al., 1999; see also Rathore et al., 2000), for example, physicians viewed video tapes of White and Black actors of both genders playing the role of patients complaining about chest pains.[2] Physicians were less likely to refer the African American than the White patients for further testing; this was especially true for African American female patients.

Many African Americans are aware of this disparity in treatment; thus, they are more likely than Whites to endorse beliefs that the medical field is biased against their group (Boulware et al., 2003). Furthermore, the more African Americans and Latino Americans endorse these beliefs, the more likely they are to prefer a same-race physician (Chen et al., 2005). In addition, the more discrimination previously experienced in her daily life, the less likely an African American patient is to subsequently adhere to a non–African American physician's recommendation (Penner et al., 2009b). Despite patients' beliefs and preferences, which could be biased in their own right, might physicians interact with and treat actual patients differently depending on the patient's race? In this section, we focus on how medical interactions between White physicians and ethnic minority patients might

be plagued by racial bias and thus contribute to racial disparities in health care and outcomes.

Medical interactions tend to meet all but one of the primary conditions posited by Allport (1954) as the ideal factors to foster harmonious intergroup encounters. The one missing piece is that health-care providers and patients are not of equal status—health-care providers, especially physicians, have higher status in this context than the patient. However, Allport's other three conditions are generally met. Health-care providers and patients have the common goal of making the patient healthier, which means they are likely to be willing to work together. Moreover, they are interdependent: the health-care provider's actions have direct effects on the patient, and the patient's decision to follow the health-care provider's advice influence the health-care provider's success rate. Finally, the American Medical Association and the federal government support and encourage medical interactions across the racial divide, often providing financial resources for White health-care providers to work in predominately ethnic minority communities. Thus, three of the four primary conditions are satisfied for successful intergroup interaction. But, are interactions between White health-care providers and ethnic minority patients of high quality, or do they suffer as a result of intergroup bias? More important, might policies be implemented to make these interactions more successful, ultimately reducing care disparities between Whites and African Americans?

As we noted previously, racial bias has become more subtle over time, in part because of changes in social norms and laws. This becomes clear when one looks at Whites' behaviors toward ethnic minorities, compared to other Whites, when the norms for appropriate behavior are ambiguous and uncertain (Dovidio and Gaertner, 2004). In these instances, Whites tend to be biased against ethnic minorities but justify their behavior to factors other than race. This is evident in the medical area, for example, when standard protocol is ambiguous and there is high level of clinical uncertainty. In such cases, health-care providers are left to make decisions using their own discretion, under cognitive demand because of time pressure and resource constraints (Smedley, Stith, and Nelson, 2003). Indeed, evidence suggests that it is under these circumstances that medical-care providers' biases come into play the most (see Penner et al., 2007, for a review).

Subtle racial bias is also apparent in the verbal and nonverbal behaviors displayed by health-care providers during interactions with patients. In general, health-care providers display more negative and less supportive behaviors toward ethnic minority, compared to White, patients. For example, physicians are less patient-centered, more verbally dominant, and express lower levels of positive affect with African American, compared to White, patients (Johnson et al., 2004). In addition, when African American patients seek information from their physicians, they receive a lower proportion of information in return compared to White patients (Gordon et al., 2006). These differences in behaviors are most pronounced in interactions when the physician is White and the patient is ethnic minority, though all interracial medical interactions (i.e., even African American physicians and White patients) are of poorer quality compared to same-race ones (Ferguson and Candib, 2002; Saha et al., 1999). Interracial medical interactions are shorter, involve slower speech, and are characterized by less positive patient affect, as assessed by independent coders of the interaction, compared to same-race medical interactions, even after controlling for variables that may influence these outcomes, such as age and health status (Cooper et al., 2003). Moreover, White physicians are less likely to provide sufficient medical information and are less likely to encourage African American patients to participate in medical decision making compared to White patients (Cooper-Patrick et al., 1999). Differences in behaviors toward Whites and ethnic minorities are also evident among newly established health-care providers. For example, White medical students and residents have significantly decreased rapport with Hispanic patients (Hooper et al., 1982; Shapiro and Saltzer, 1981), and display fewer positive expressions in their speech to Hispanic, compared to White, patients (Sleath, Rubin, and Arrey-Wastavino, 2000). In sum, accumulating evidence suggests that ethnic minority patients are the target of more negative interpersonal treatment during medical interactions, especially with White health-care providers (for reviews see Dovidio et al., 2008; Penner et al., 2007; van Ryan and Fu, 2003).

Differences in the behaviors of White health-care providers during interactions with ethnic minority, compared to White, patients do not occur without costs. Many of the negative communication patterns noted above are associated with lower patient compliance, patient satisfaction, and health-care outcomes (Stewart, 1995). For example, when doctors are dominant and less informative, patients are less likely to have a strong grasp on their health-treatment options and less likely to comply with the options they are given (Hall, Roter, and Katz, 1988). Therefore, interventions and policies are needed in order to combat these problems in interracial interactions in the medical context.

The research described thus far implies that racial bias exists in medical interactions because of the differences in behaviors based on the patient's race. Ad-

ditional research, albeit limited, provides direct evidence that the racial beliefs of health-care providers impact their behaviors toward patients. In a sample of predominately White health-care providers, for example, physicians rated African American patients as less educated and intelligent, abusers of drugs, less desiring of a physically active lifestyle, and less compliant with medical advice (van Ryn and Burke, 2000; van Ryn et al., 2006). Endorsement of these beliefs, in turn, influenced physicians' recommendations for treatment. Specifically, perceptions of patients' education and physical activity levels were responsible for African American patients being recommended for bypass surgery less often than White patients (van Ryn et al., 2006). In similar work, Green et al. (2007) found that the implicit racial attitudes of White health-care providers predicted their behaviors toward minority patients.

Consistent with the U.S. Department of Health and Human Services' goal to eliminate racial disparities in health, medical institutions, such as the Institute of Medicine, have put forth various policy recommendations for ways to reduce racial biases in medical contexts (Betancourt and Maina, 2004). We describe three of these policy recommendations here. One recommendation is to create policies that strengthen physician-patient relationships in publicly funded health plans (Smedley, Stith, and Nelson, 2003). As we explained earlier, the psychological evidence demonstrates that people who are highly motivated to be nonprejudiced can control even the subtle effects of implicit biases if they have the time and cognitive resources to do so. One direct policy implication of these findings is that guidelines should be created and enforced that limit patient loads per primary physicians in order to reduce the psychological and material resource constraints that make it easy for physicians' biases to interfere with the medical interaction. Similarly, according to the Institute of Medicine, policies are needed that enforce time allotments for patient visits (longer times allotted and allowance for more time when necessary) to facilitate smoother interactions.

A second recommendation by the Institute of Medicine is to integrate cross-cultural education into the training of all health-care professionals. Indeed, the Liaison Committee on Medical Education and the Accreditation Council on Graduate Medical Education strongly encourage that cross-cultural curricula be integrated into medical education. Although such practice is encouraged, it is not always fulfilled (Betancourt, 2006; Welch, 1998). When cultural competence training does occur, students are made aware of sociocultural influences on health beliefs and behaviors. Moreover, they are taught how biases develop from normal categorization processes and can influence medical decisions. Indeed, cultural sensitivity training has been shown to increase health-care providers' intergroup tolerance and openness to others (Culhane-Pera et al., 1997). In addition, health-care providers with cross-cultural educational training have shown improvements in interpreting verbal and nonverbal cues communicated to them by out-group members (Majumdar, 1999). Furthermore, simply having medical students reenact experiences of perceived discrimination or insensitive behaviors improves their cultural sensitivity (Johnston, 1992).

Unfortunately, however, one of the limitations of many of the existing cross-cultural training programs is that the programs focus on changes in participants' intergroup attitudes as a result of being in the program instead of the effectiveness of the training on participants' actual behaviors in medical interactions. Although changing health-care providers' attitudes is essential, it is also crucial to assess whether or not these changes translate into improvements in behavior. There is indirect evidence, however, that cross-cultural training improves health-care providers' behaviors toward their patients. Specifically, Majumdar et al. (2004) found that patients who were treated by health-care providers with cross-cultural training reported overall better functioning than patients who were treated by health-care providers without such training.

Finally, the Institute of Medicine committee recommended training more ethnic minority, especially African American, health professionals. Among medical school graduates in 2001, only 10% were ethnic minorities. Therefore, there is a dire need to enforce affirmative action policies in medical school admissions and residency recruitment in order to ensure a diverse medical profession. This would allow minority patients to have a higher probability of having race-concordant medical interactions. In addition, White physicians' racial attitudes about ethnic minorities may change as a result of interacting with other in-group members (physicians) who differ from them only in terms of race.

In addition to the recommendations by the Institute of Medicine, we offer several policy recommendations that focus specifically on improving the dynamics of medical care provider–patient interactions as a means to reduce health disparities. Similar to Penner et al. (2007), one practice we recommend is for policy makers to emphasize continuity of care, meaning that a patient sees the same provider or small group of providers who interact with one another over an extended period of time. Given that regular contact with members of different racial groups reduces anxiety and stereotyping (Pettigrew and Tropp,

2006), White health-care providers are likely to become more comfortable with the same ethnic minority patient over time, perceiving them in more individualized (decategorized) ways and decreasing the chance for unintended biases to be problematic in providing excellent care to their patient. Moreover, contact with the same ethnic minority patient over time is likely to increase rapport and empathy as medical providers learn more personal information about their patients. Furthermore, patients are likely to become more trusting of their medical-care providers; trust, for good reasons, tends to be low among African American patients (Halbert et al., 2006). In settings in which it is difficult to ensure that a patient will see the same provider, such as walk-in clinics that have limited permanent staff, it is often possible to create patient-provider "teams," which can enhance a sense of psychological connection as well as offer continuity in medical care (Penner et al., 2009a).

Even when patients regularly see the same provider, creating perceptions of a physician-patient team can reduce bias and improve the quality of interaction in medical encounters between physicians and patients of different races through the principles outlined in the Common Ingroup Identity Model (Gaertner and Dovidio, 2000). That is, the goal is to encourage medical-care providers to view themselves and their patients as members on the same team instead of viewing themselves as members of one group (medical-care providers) and their patients as members of a different group (people needing help). This approach is likely to foster a sense that the medical-care provider and patient are collaborators who are working together to make the best medical decision.

Another practice we recommend is to provide a safe space, perhaps as part of a diversity training workshop, for medical-care providers to be made aware of their potential for racial bias during interactions with patients. Because of their egalitarian values it is common for people to deny that they are biased and that their behaviors may reflect racial bias. This is likely because people tend to focus on blatant forms of bias, and in contemporary American society, expressing blatant racial bias during interactions, including interactions between health-care providers and patients (Epstein, 2005), is not common. In fact, research shows that most medical providers deny that a patient's race plays any role in their medical decisions (Lurie et al., 2005), though it is clear that it does (Dovidio et al., 2008). Research shows that when low-prejudiced people are made aware of the discrepancy between their personal beliefs and actual behavior toward racial out-group members, they feel compunction and are motivated to change their behavior in the future (Monteith and Mark, 2005). If medical-care providers are made aware that their behaviors toward patients not of their racial group are biased, they are likely to try to change those behaviors, and research suggests that with practice they will be able to do so (Dovidio, Kawakami, and Gaertner, 2000; Kawakami et al., 2000). Based on research revealing that increased knowledge of another group has only limited effects for improving intergroup relations (Pettigrew and Tropp, 2008), Burgess et al. (2007) proposed that programs for medical students not only increase awareness of the complex nature of contemporary intergroup bias but also instruct and train students in the skills (e.g., partnership building with patients and emotion regulation) that can help them control even implicit biases in their medical interactions.

In this section we have suggested the different ways that intergroup biases can shape medical encounters in ways that can contribute, often without the parties' awareness or intention, to racial disparities in health care and, ultimately, health. In the next section, we illustrate the operation of these processes in a very different context—the workplace.

Workplace Context

The Department of Labor estimated in 2007 that the average employee spends approximately 1,900 hours a year working, which is equivalent to about one-third of one's waking hours. Given that the number of ethnic minorities in the workforce has increased as a result of federal laws, specifically Title VII of the Civil Rights Act of 1964, that not only prohibit discrimination against hiring minorities but also encourage action to hire minorities, many of these working hours are likely to involve interactions across racial lines. With respect to the conditions Allport (1954) posited as facilitating harmonious intergroup interactions, all but one is usually met in the workplace. Although there are certainly cases where people of different racial groups but of equal status interact in the workplace, given the racial segregation of jobs in the United States (Forman, 2003), it is more likely the case that Whites are in the higher-status role, whereas ethnic minorities are in the lower-status role (i.e., White supervisor, African American subordinate). The other conditions, however, are generally met.

Despite federal laws promoting racial equality, there is widespread evidence that racial disparities and discrimination still persist in employment processes, from recruitment to advancement as well as in typical everyday interactions in organizations (Brief, Butz, and Deitch, 2005; see Fiske and Krieger, this volume, for policy issues surrounding gender discrimination). As we have shown with similar findings among students in residential settings on college campuses and

health-care providers and patients in medical settings, people prefer to interact with other racial in-group members in the workplace (Williams and O'Reilly, 1998). This is the case for coworkers working together on various business tasks as well as for mentor-protégé relationships. White managers, for example, are less likely to form mentoring relationships with ethnic minorities than with White protégés (Thomas, 1993). This disparity is quite unfortunate because mentoring relationships are wonderful resources for promotions and career mobility. Given that there are not often many ethnic minorities in higher-status positions in corporations, ethnic minorities may not receive the mentoring they need for upward career mobility; thus, the status quo regarding race and leadership roles in companies is enforced.

In addition to a general preference for working with in-group members in the workplace, employees tend to be biased against out-group members even when important qualifications, such as skills, are equivalent across different group members. For example, "audit studies" in which researchers send research assistants into the field as potential applicants have shown different treatment of African American and White job applicants who were matched in qualifications and interviewing skills (e.g., Pager and Western, 2006). Specifically, African American applicants are less likely to receive an interview and eventually be hired than White applicants. However, even if ethnic minority job applicants are just as likely as White job applicants to receive an interview, their experience during the process of inquiring about a job is often less positive than that of Whites, especially if their ethnicity is made salient to the employee. In an illustration of this, Barron, Hebl, and King (2011) had White and ethnic minority research assistants pose as potential job applicants. The applicants entered different stores and asked the manager on duty for a job application. During this process, the assistants wore hats that either mentioned their ethnicity (e.g., Black Student Association; Asian American Student Association) or had no reference to ethnicity (i.e., Rice Student Association), but the students were unaware of the message on the hat. After interacting with the manager, the assistants rated their experience, and independent coders rated the quality of the interaction. Results revealed that employers were not biased against the ethnic minority applicants on formal employment behaviors, such as permission to complete a job application and callbacks for further consideration. However, bias was expressed more subtly in the employers' interaction behaviors. The employers spent less time and used fewer words when interacting with the ethnic minority applicants whose race was made salient by the hats than their White counterparts. Taken together, these findings reveal that ethnic minority job applicants receive differential treatment during interactions from the initial stages of the employment process, which may undermine their desire to pursue employment in a company.

Moreover, if ethnic minorities do follow through with an interview, bias is apt to come into play during the interview, making it difficult for them to have a successful interview and ultimately be hired. As we have noted before, racial bias is most likely to occur in less structured situations and when nonracial explanations can be used to justify individuals' decisions (Brief et al., 2000; Elvira and Zatzick, 2002; Huffcut and Roth, 1998). In the workplace this means that when the interview is less structured, evaluations of African American and Hispanic applicants during the employment interview are less favorable than those of White applicants (Huffcut and Roth, 1998). For example, Dovidio and Gaertner (2000) illustrated that in a simulated hiring situation, Whites were not biased against African American job applicants compared to White job applicants when the applicant's qualifications were clearly weak or strong. However, when the job applicant's qualifications were moderate, which creates an ambiguous situation, Whites were biased against the African American candidate because they were able to justify their decision on a factor other than race (i.e., lack of skills). Moreover, Son Hing et al. (2008) further demonstrated that racial bias in hiring for moderately qualified candidates is predicted by people's implicit racial bias. Thus, employers may not be aware of how their biases shape their perceptions of minority candidates, ultimately limiting the latters' employment opportunities.

Moreover, racial biases are often leaked through nonverbal behaviors during interactions, which can interfere with a successful interview process. Word, Zanna, and Cooper (1974) illustrated this by having naive White interviewers interview White and African American interviewees, who were trained to behave in a similar fashion. They found that the naive White interviewers displayed less friendly nonverbal behaviors toward African American interviewees than toward White interviewees. Specifically, the White interviewers were more physically distant, asked fewer questions, and made less eye contact during interviews with African American, compared to White, interviewees. In a second study, Word, Zanna, and Cooper (1974) trained White interviewers (confederates) to display either friendly or unfriendly nonverbal behaviors toward naive White interviewees. The results revealed that the naive, White interviewees who were the target of unfriendly nonverbal behaviors performed worse during the interview than those who were the target of friendly nonverbal behaviors.

Thus, the White interviewees who received the same sort of unfriendly treatment that the African American interviewees in the first study had received did not perform well during the interview. Taken together, these findings show that unstructured interviews and differences in nonverbal behaviors during structured interviews can be detrimental to the career advancement of ethnic minorities.

Not only do White employers' personal biases influence their behaviors during interactions with job applicants and employees, but biases intrinsic to the workforce may influence employers' behaviors, ultimately affecting ethnic minorities' success in business. Moreover, ethnic minorities' sensitivity to Whites' nonverbal behaviors may also create disadvantages for ethnic minorities' performance in the workplace. These two issues were illustrated in a recent study on behavioral mimicry in interethnic interactions in the workplace (Sanchez-Burks, Blount, and Bartel, 2007). *Behavioral mimicry* is the process by which people unknowingly change the timing and content of their behaviors so that they mirror the behavioral cues expressed by their partner (e.g., Chartrand and Bargh, 1999). People tend to feel more comfortable with and like those who mirror them than those who do not, and they tend to mirror people they feel comfortable with and like (Chartrand and Bargh, 1999; Lakin and Chartrand, 2003). People are generally unaware that they are engaging in behavioral mimicry, and when they are aware that their interaction partner is mimicking them, they show less rapport with that partner than when they are unaware of this behavior (Chartrand, Maddux, and Lakin, 2005).

Perhaps because in the American workplace more attention is placed on task concerns than on interpersonal relationships and emotions, Sanchez-Burks, Blount, and Bartel (2007) suggested that White Americans are less likely to engage in behavioral mimicry when interacting with subordinates, regardless of the subordinates' cultural background. Because of their cultural traditions, Latinos are more sensitive to relational cues, such as behavioral mimicry, than Whites (Sanchez-Burks, Bartel, and Blount, 2009). As a result, interactions between a White manager who does not engage in behavioral mimicry and a Latino subordinate may be less positive than interactions between this same White manager and a White subordinate. To test this idea, Sanchez-Burks, Blount, and Bartel (2007) had White and Latino midlevel employees of a United States Fortune 500 company participate in a mock interview in a headquarters' office suite. Participants were assigned to be interviewed by a White interviewer who mimicked or did not mimic the employees' behaviors in a subtle manner. The presence of behavioral mimicry decreased the amount of anxiety among Latinos but not among Whites. Moreover, the presence of behavioral mimicry improved Latinos' self-reported performance evaluations as well the experts' ratings of their performance, but this did not occur with White employees. Thus, the subtle nonverbal behaviors (or lack thereof), and especially the attention to these behaviors (or attention to the lack of such behaviors), has the potential to have a greater impact on ethnic minorities' performances in the workplace than on Whites'. This is quite unfortunate because ethnic minorities are most often in subordinate positions in the workplace, and these findings suggest that their continued awkward, uncoordinated interactions with White superiors may interfere with their advancement in business.

Given that a diverse workplace has the potential to promote creativity as well as improve employees' individual and group performances (Ely and Thomas, 2001), policies that are designed to alleviate impediments to harmonious intergroup interactions are essential to the success of a company. One policy that some companies have wrestled with is diversity training as a means to improve intergroup relations among employees. There are many kinds of diversity training programs, but they are not always successful at their ultimate goal of decreasing intergroup bias and improving intergroup interactions. Arthur and Doverspike (2005) summarized the most important elements needed for these programs to be effective. Specifically, (a) the training must emphasize dispelling stereotypes instead of avoiding them; (b) specific steps must be included to show workers how to translate their positive racial attitudes into positive behaviors; (c) there must be sufficient time available for training; and (d) the training must be sanctioned by top management, but it should not come across as occurring merely for workers to be politically correct in their behavior.

In addition to those elements, we recommend that diversity training programs focus on fostering a common in-group identity among employees of diverse backgrounds. That is, employees should be encouraged to think of themselves as a superordinate group (i.e., employees of Company X) who are working together to produce the best outcome in the most efficient manner instead of individuals from different racial groups. Indeed, research shows that Whites who were induced to perceive themselves as teammates with an African American partner evaluated their partner more positively than Whites who perceived themselves as individuals who were just working on the same task as an African American (Nier et al., 2001). Furthermore, banking executives who had experienced a corporate merger with various other banks had more favorable attitudes toward the executives of the various banks

when a common identity group was created (Bachman, 1993, as cited in Dovidio and Gaertner, 2004). Taken together, it is likely that diversity training programs that foster a common identity among employees of diverse backgrounds will improve the dynamics of interactions among these employees.

Conclusion: Intergroup Dynamics and Policy: Common Threads

In this chapter, we described the complex nature of intergroup attitudes, which often includes unconscious prejudice that is manifested in subtly biased ways rather than in terms of blatant discrimination. Moreover, we have outlined the particular challenges of interracial interactions, in which the different perspectives and expectations of Whites and African Americans can produce miscommunication and misunderstandings, and often divergent perspectives that reinforce intergroup mistrust. We have also shown that across three contexts—residential spaces on college campuses, medical facilities, and workplaces—people prefer and have more positive experiences during same-race, rather than interracial, interactions, make decisions that favor in-group members, and behave more positively toward in-group members. Their negative decisions and behaviors, however, are not always blatant and generally occur when the situation is less structured so that it is more ambiguous as to whether or not race played a role. We have also discussed policies used in each setting to combat interracial bias. Clearly, policies and interventions need to be tailored to particular issues in specific contexts. The influence of intergroup biases is significantly different for college roommates, in medical provider-patient interactions, and in employer hiring decisions. Nevertheless, there are common threads in the dynamics of bias across these situations and thus similar challenges for policies and interventions to improve the different outcomes.

One common feature of these different situations is that bias is expressed in subtle ways: in more complex and thus potentially strained relationships between roommates, shorter and less effective interactions between physicians and patients, and in lower levels of rapport in employment interviews. In each of these situations, there is rarely an obvious, single action that qualifies as blatant discrimination. Thus, traditional policies that were designed to respond to overt forms of discrimination, although still important today, are not sufficient for addressing the forms of bias that may be contributing to social problems in each of the areas we discussed. The contemporary challenge for policies may be to promote ways that intergroup relations can be as positive and productive as relations among members of the same racial or ethnic group.

There are two common practices that policy makers have taken in attempt to improve intergroup relations across the three contexts we have explored in this chapter. First, it is clear that action has been taken to implement ways to make sure that there is the opportunity for intergroup contact to occur. This means not only increasing the number of ethnic minorities in these settings, but also making sure that members of the diverse group actually interact with one another. For example, most universities have an explicit rule that race cannot be used as a factor to cluster students of the same racial group together in residential spaces. However, these same universities also have an implicit rule that race should be used to diversify these same spaces. Similarly, many medical administrators highly recommend that medical schools not only admit more ethnic minority applicants, but also encourage that all students become involved in training that allows them to interact with a diverse population, such as working in hospitals that are in a predominately ethnic minority community. And, many companies and businesses attempt to increase their number of ethnic employees as well as make sure that their work teams are diverse, requiring individuals to interact with out-group members. Mere contact, as Allport (1954) noted, is not enough. Thus, policy makers need to move beyond the first step of opening the doors to a diverse group.

As a second step toward moving beyond simply increasing the number of people from diverse backgrounds, another practice that has been used by policy makers to improve intergroup encounters is multicultural training education programs that enhance knowledge of other cultures and acceptance of a diverse society. Although the content is not exclusive to residential spaces, many universities require students to attend sessions about diversity as a part of their freshmen orientation. Likewise, the Institute of Medicine recommends that cross-cultural education be a required part of medical training. Indeed, some medical schools have followed this advice and made diversity training classes required, whereas others have included these classes in the curriculum only as electives. And, perhaps hoping for a potential financial return, many companies and organizations also offer diversity training workshops for their employees.

Although diversity training is a step beyond simply increasing the number of minorities in a community, it is important to note that the research on the effectiveness of multicultural training has been, at best, mixed. Most of this research suggests that this type of training is good for changing explicit racial attitudes

but may be less effective for changing implicit racial attitudes and nonverbal behaviors, which is a major means by which unconscious bias is expressed. Throughout this chapter, we have focused on findings from social psychological research that we believe would be most useful to policy makers as they develop these diversity training programs. We emphasize, for example, that members of the diversity training workshop should be made aware of their potential for intergroup bias. In addition, we recommend creating a common identity among the different groups instead of focusing on separate racial groups. These factors have been shown to facilitate harmonious interracial interactions and reduce racial bias. Thus, policy initiatives that include these practices are likely to be quite promising in changing race relations in American society.

In conclusion, we identify three fundamental principles for guiding policies and interventions across different contexts. The first principle that forms the basis of all our recommendations is that diversity needs to be acknowledged even while recognizing common connections and shared identities across group lines. Put simply, the effects of racism, which involve the perceptions and actions of Whites and African Americans, among others, cannot be addressed through policies that ignore race as a determining factor (e.g., color-blind policies). Indeed, research has found that multicultural perspectives on racial diversity are associated with less racial bias than are colorblind perspectives (Richeson and Nussbaum, 2004; Wolsko, Park, and Judd, 2006). Moreover, Plaut and colleagues found that racial minorities who worked in departments with White co-workers who endorsed a multicultural ideology were more engaged in their departments than were racial minorities who worked in departments with Whites who endorsed colorblindness (Plaut, Thomas, and Goren, 2009). In other words, the endorsement of multiculturalism was associated with greater feelings of belonging and engagement among minority employees. Although color blindness may have unexpected negative outcomes, the acknowledgement of group memberships should not reinforce race as a biological concept but, simply, recognize it as a social reality in contemporary American society.

The second common principle underlying all our recommendations is that for policies and interventions to be effective, they have to consider the complex and often subtle nature of intergroup bias today. As we noted in our review of the literature on interracial interaction, Whites and Blacks often have different interpretations of and responses to interracial contact situations, and interventions need to address the needs of members of the different groups. Because intergroup *relations* and outcomes are shaped by members of the different groups, policy makers need to understand those different perspectives and address the unique needs, as well as the common objectives, of the members of the groups.

And finally, our third principle is that policies need to consider both the long-term and short-term effects, weighing both with an eye to achieving the ultimate goal of the intervention. For example, the research on university residences revealed that relationships between roommates of different races are often more strained than are relationships between roommates of the same race. Nevertheless, having a roommate of a different race can, over time, produce more positive intergroup attitudes and relations on campus. In other contexts, groups with racial and ethnic diversity often experience more social tensions than do racially or ethnically homogeneous groups (Putnam, 2007), but at the same time, diverse groups are better at solving complex problems that require divergent thinking (Antonio et al., 2004) and attending to a broader range of relevant information in the analysis of issues (Sommers, 2006). Policies aimed at achieving immediate harmony may thus preclude achieving other, often more desirable, long-term benefits of diversity.

Notes

1. The committee created to examine the role of intergroup biases in racial disparities in health outcomes "found no direct evidence that racism, bias, or prejudice among healthcare professionals affects the quality of care for minority patients, such as that which might be available from audit studies where 'testers' from different racial or ethnic groups present in clinical settings with similar clinical complaints, histories, and symptoms to assess possible differences in the quality of their treatment" (Smedley, Stith, and Nelson, 2003, p. 176). However, the committee acknowledged that bias, stereotyping, and prejudice on the part of health care providers can not be ruled out.

2. Researchers who reanalyzed the dataset claim Schulman's et al. findings were incorrect (e.g., Schwartz, Woloshin, and Welch, 1999).

References

Allport, G. W. (1954). *The nature of prejudice*. Reading, MA: Addison-Wesley.

Antonio, A. L., Chang, M. J., Hakuta, K., Kenny, D. A., Levin, S., and Milem, J. F. (2004). Effects of racial diversity on complex thinking in college students. *Psychological Science*, *15*, 507–510.

Aron, A., Eberhardt, J. L., Davies, K., Wright, S. C., and Bergsieker, H. B. (2007). *A social psychological inter-*

vention to improve community relations with police: Initial results. Manuscript in progress.

Aron, A., Melinat, E., Aron, E. N., Vallone, R. D. and Bator, E. (1997). The experimental generation of interpersonal closeness: A procedure and some preliminary findings. *Personality and Social Psychology Bulletin*, *23*, 363–377.

Arthur, W., Jr., and Doverspike, D. (2005). Achieving diversity and reducing discrimination in the workplace through human resource management practices: Implications of research and theory for staffing, training, and rewarding performance. In R. L. Dipboye and A. Colella (Eds.), *Discrimination at work* (pp. 305–328). Mahwah, NJ: Lawrence Erlbaum.

Associated Press. (2006, March 12). Officials seek to limit self-segregated dorms at UMass-Amherst. *Boston.com News.* Retrieved from http://www.boston.com/news/education/higher/articles/2006/03/12/officials_seek_to_limit_self_segreated_dorms_at_umass_amherst

Barron, L. G., Hebl, M., and King, E. B. (2011). Effects of manifest ethnic identification on employment discrimination. *Cultural Diversity and Ethnic Minority Psychology*, *17*(1), 23–30.

Betancourt, J. R. (2006). Cultural competence and medical education: Many names, many perspectives, one goal. *Academic Medicine*, *81*, 499–501.

Betancourt, J., and Maina, A. (2004). The Institute of Medicine report "Unequal Treatment": Implications for academic health centers. *Mount Sinai Journal of Medicine*, *71*(5), 314–321.

Bobo, L. (2001). Racial attitudes and relations at the close of the twentieth century. In N. J. Smelser, W. J. Wilson, and F. M. Mitchell (Eds.), *Racial trends and their consequences* (Vol. 1, pp. 264–307). Washington, DC: National Academy Press.

Bodenhausen, G. V., Todd, A. R., and Becker, A. P. (2007). Categorizing the social world: Affect, motivation, and self-regulation. In B. H. Ross and A. B. Markman (Eds.), *The psychology of learning and motivation* (Vol. 47, pp. 123–155). Amsterdam: Elsevier.

Boulware, L. E., Cooper, L. A., Ratner, L. E., LaVeist, T. A., and Powe, N. R. (2003). Race and trust in the health care system. *Public Health Reports*, *118*, 358–365.

Brief, A. P., Butz, R. M., and Deitch, E. A. (2005). Organizations as reflections of their environments: The case of race composition. In R. L. Dipboye and A. Colella (Eds.), *Discrimination at work: The psychological and organizational biases* (pp. 119–148). Mahwah, NJ: Lawrence Erlbaum.

Brief, A. P., Dietz, J., Cohen, R. R., Pugh, S. D., and Vaslow, J. B. (2000). Just doing business: Modern racism and obedience to authority as explanation for employment discrimination. *Organizational Behavior and Human Decision Processes*, *81*, 72–97.

Burgess, D., van Ryn, M., Dovidio, J. F., and Saha, S. (2007). Reducing racial bias among health care providers: Lessons from social cognitive psychology. *Journal of General Internal Medicine*, *22*, 882–887.

Chartrand, T. L., and Bargh, J. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology*, *76*, 893–910.

Chartrand, T. L., Maddux, W., and Lakin, J. (2005). Beyond the perception behavior link: The ubiquitous utility and motivational moderators of nonconscious mimicry. In R. Hassim, J. Uleman, and J. Bargh (Eds.), *The new unconscious* (pp. 334–361). New York: Oxford University Press.

Chen, F. M., Fryer, G. E., Phillips, R. L., Wilson, E., and Patham, D. E. (2005). Patients' beliefs about racism, preferences for physician race, and satisfaction with care. *Annals of Family Medicine*, *3*, 138–143.

Cooper, L. A., Roter, D. L., Johnson, R. L., Ford, D. E., Steinwachs, D. M., and Powe, N. R. (2003). Patient-centered communication, ratings of care, and concordance of patient and physician race. *Annals of Internal Medicine*, *139*, 907–915.

Cooper-Patrick, L., Gallo, J. J., Gonzales, J. J., Vu, H. T., Powe, N. R., Nelson, C., and Ford, D. E. (1999). Race, gender, and partnership in the patient-physician relationship. *Journal of the American Medical Association*, *282*, 583–589.

Crandall, C. S., and Eshleman, A. (2003). A justification-suppression of the expression and experience of prejudice. *Psychological Bulletin*, *129*, 414–446.

Culhane-Pera, K. A., Reif, C., Egli, E., Baker, N. J., and Kassekert, R. (1997). A curriculum for multicultural education in family medicine. *Family Medicine*, *29*, 719–723.

Deutsch, M., and Collins, M. E. (1951). *Interracial housing; A psychological evaluation of a social experiment.* Minneapolis, MN: University of Minnesota Press.

Dovidio, J. F. (2001). On the nature of contemporary prejudice: The third wave. *Journal of Social Issues*, *57*, 829–849.

Dovidio, J. F., and Gaertner, S. L. (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological Science*, *11*, 319–323.

———. (2004). Aversive racism. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 36, pp. 1–52). San Diego, CA: Elsevier Academic Press.

Dovidio, J. F., Kawakami, K., and Beach, K. (2001). Implicit and explicit attitudes: Examination of the relationship between measures of intergroup bias. In R. Brown and S. L. Gaertner (Eds.), *Blackwell handbook of social psychology: Intergroup relations* (Vol. 4, pp. 175–197). Oxford, UK: Blackwell.

Dovidio, J. F., Kawakami, K., and Gaertner, S. L. (2000). Reducing contemporary prejudice: Combating explicit

and implicit bias at the individual and intergroup level. In S. Oskamp (Ed.), *Reducing prejudice and discrimination* (pp. 137–163). Mahwah, NJ: Lawrence Erlbaum.

———. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, *82*, 62–68.

Dovidio, J. F., Penner, L. A., Albrecth, T. L., Norton, W. E., Gaertner, S. L., and Shelton, J. N. (2008). Disparities and distrust: The implications of psychological processes for understanding racial disparities in health and health care. *Social Science and Medicine*, *67*, 478–486.

Dunton, B. C., and Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin*, *23*, 316–326.

Elvira, M. M., and Zatzick, C. D. (2002). Who's displaced first? The role of race in layoff decisions. *Industrial Relations*, *41*, 69–85.

Ely, R., and Thomas, D. (2001). Cultural diversity at work: The effects of diversity perspectives on work group processes and outcomes. *Administrative Science Quarterly*, *46*, 229–274.

Epstein, R. A. (2005). Disparities and discrimination in health care coverage: A critique of the Institute of Medicine study. *Perspectives in Biology and Medicine*, *48*, S26–S41.

Fazio, R. H. (1990). A practical guide to the use of response latency in social psychological research. In C. Hendrick and M. Clark (Eds.), *Research methods in personality and social psychology* (pp. 74–97). Thousand Oaks, CA: Sage Publications.

Ferguson, W. J., and Candib, L. M. (2002). Culture, language, and the doctor-patient relationship. *Family Medicine*, *34*, 353–361.

Forman, T. A. (2003). The social psychological costs of racial segmentation in the workplace: A study of African Americans' well-being. *Journal of Health and Social Behavior*, *44*, 332–352.

Gaertner, J. F. and Dovidio, S. L. (1986). Prejudice, discrimination, and racism: Historical trends and contemporary approaches. In S. Gaertner and J. Dovidio (Eds.), *Prejudice, discrimination, and racism* (pp. 1–34). San Diego, CA: Academic Press.

———. (2000). Reducing intergroup bias: The common ingroup identity model. *Reducing intergroup bias: The common ingroup identity model*. New York: Psychology Press.

Gordon, H. S., Street, R. L., Sharf, B. F., and Soucheck, J. (2006). Racial differences in doctors' information-giving and patients' participation. *Cancer*, *107*, 1313–1320.

Green, A. R., Carney, D. R., Pallin, D. J., Ngo, L. H., Raymond, K. L., Iezzoni, L. I., and Banaji, M. R. (2007). The presence of implicit bias in physicians and its predictions of thrombolysis decisions for Black and White patients. *Journal of General Internal Medicine*, *22*, 1231–1238.

Halbert, C. H., Armstrong, K., Gandy, O. H., and Shaker, L. (2006). Racial differences in trust in health care providers. *Archives of Internal Medicine*, *166*, 896–901.

Hall, J. A., Roter, D. L., and Katz, N. R. (1988). Meta-analysis of correlates of provider behavior in medical encounters. *Medical Care*, *26*, 657–675.

Hebl, M. R., and Dovidio, J. F. (2005). Promoting the "social" in the examination of social stigmas. *Personality and Social Psychology Review*, *9*, 156–182.

Hebl, M. R., and King, E. (2007). *Reducing interpersonal discrimination*. Paper presented at the Small Group Meeting: Social Stigma and Social Disadvantage. Leiden University, the Netherlands.

Hooper, E. M., Comstock, L. M., Goodwin, J. M., and Goodwin, J. S. (1982). Patient characteristics that influence physician behavior. *Medical Care*, *20*, 630–638.

Huffcut, A. I., and Roth, P. L. (1998). Racial group differences in employment interview evaluations. *Journal of Applied Psychology*, *83*, 179–189.

Ickes, W. (1984). Compositions in Black and White: Determinants of interaction in interracial dyads. *Journal of Personality and Social Psychology*, *47*, 330–341.

Johnson, R. L., Roter, D., Powe, N. R., and Cooper, L. A. (2004). Patient race/ethnicity and quality of patient-physician communication during medical visits. *American Journal of Public Health*, *94*, 2084–2090.

Johnston, M. A. (1992). A model program to address insensitive behaviors toward medical students. *Academic Medicine*, *67*, 236–237.

Kawakami, K., Dovidio, J. F., Moll, J., Hermsen, S., and Russin, A. (2000). Just say no (to stereotyping): Effect of training in the negation of stereotypic associations on stereotype activation. *Journal of Personality and Social Psychology*, *78*, 871–888.

Lakin, J. L., and Chartrand, T. L. (2003). Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychological Science*, *14*, 334–340.

Livingston, R. W. (2002). The role of perceived negativity in the moderation of African Americans' implicit and explicit racial attitudes. *Journal of Experimental Social Psychology*, *38*(4), 405–413. doi:10.1016/S0022-1031(02)00002-1

Lurie, N., Fremont, A., Jain, A. K., Taylor, S. L., McLaughlin, R., Peterson, E., et al. (2005). Racial and ethnic disparities in care: The perspectives of cardiologists. *Circulation*, *111*, 1264–1269.

Majumdar, B., Keystone, J. S., and Cuttress, L. A., (1999). Cultural sensitivity training among foreign medical graduates. *Medical Education*, *33*, 177–184.

Majumdar, B., Browne, G., Roberts, J., and Carpio, B. (2004). Effects of cultural sensitivity training on health care provider attitudes and patient outcomes. *Journal of Nursing Scholarship*, *36*, 161–166.

Martin, W. E. (1998). *Brown v. Board of Education: A*

brief history with documents. New York: Bedford, St. Martin's Press.

Monteith, M. J., and Mark, A. Y. (2005). Changing one's prejudiced ways: Awareness, affect, and self-regulation. *European Review of Social Psychology*, *16*, 113–154.

Nier, J. A., Gaertner, S. L., Dovidio, J. F., Banker, B. S., and Ward, C. M. (2001). Changing interracial evaluations and behavior: The effects of common group identity. *Group Processes and Intergroup Relations*, *4*, 299–316.

Olson, M. A., and Fazio, R. H. (2004). Trait inferences as a function of automatically activated racial attitudes and motivation to control prejudiced reactions. *Basic and Applied Social Psychology*, *26*, 1–11.

Page-Gould, E., Mendoza-Denton, R., and Tropp, L. R. (2008). With a little help from my cross-group friend: Reducing anxiety in intergroup contexts through cross-group friendship. *Journal of Personality and Social Psychology*, *95*, 1080–1094.

Pager, D., and Western, B. (2006). *Race at work: Realities of race and criminal record in the New York City job market.* Report prepared for the 50th Anniversary of the New York City Commission on Human Rights.

Penner, L. A., Albrecht, T. L., Coleman, D. K., and Norton, W. E. (2007). Interpersonal perspectives on Black-White health disparities: Social policy implications. *Social Issues and Policy Review*, *1*, 63–98.

Penner, L. A., Dailey, R. K., Markova, T., Porcelli, J. H., Dovidio, J. F., and Gaertner, S. L. (2009a, February). *Using the common ingroup identity model to increase trust and commonality in racially discordant medical interactions.* Poster presented at the Annual Meeting of the Society for Personality and Social Psychology, Tampa, FL.

Penner, L. A., Dovidio, J. F., Edmondson, D., Dailey, R. K., Markova, T., Albrecht, T. L., and Gaertner, S. L. (2009b). The experience of discrimination and Black-White health disparities in medical care. *Journal of Black Psychology*, *35*(2), 180–203.

Pettigrew, T. F. (1997). Generalized intergroup contact effects on prejudice. *Personality and Social Psychology Bulletin*, *23*, 173–185.

Pettigrew, T., and Tropp, L. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, *90*, 751–783.

———. (2008). How does intergroup contact reduce prejudice? Meta-analytic tests of three mediators. *European Journal of Social Psychology*, *38*, 922–934.

Phelps, R. E., Altschul, D. B., Wisenbaker, J. M., Day, J. F., Cooper, D., and Potter, C. G. (1998). Roommate satisfaction and ethnic identity in mixed-race and White university roommate dyads. *Journal of College Student Development*, *39*, 194–203.

Phelps, R. E., Potter, C. G., Slavich, R., Day, J. F. and Polovin, L. B. (1996). Roommate satisfaction and ethnic pride in same-race and mixed-race university roommate pairs. *Journal of College Student Development*, *37*, 377–388.

Plant, E. A., and Devine, P. G. (2001). Responses to other-imposed pro-Black pressure: Acceptance or backlash? *Journal of Experimental Social Psychology*, *37*, 486–501.

———. (2003). The antecedents and implications of interracial anxiety. *Personality and Social Psychology Bulletin*, *29*, 790–801.

Plaut, V. C., Thomas, K. M., and Goren, M. J. (2009). Is multiculturalism or color blindness better for minorities? *Psychological Science*, *20*(4), 444–446. doi:10.1111/j.1467-9280.2009.02318.x

Purdie-Vaughns, V., Steele, C. M., Davies, P. G., Ditlmann, R., and Crosby, J. R. (2008). Social identity contingencies: How diversity cues signal threat and safety for African Americans in mainstream institutions. *Journal of Personality and Social Psychology*, *94*, 615–630.

Putnam, R. D. (2007). *E Pluribus Unum*: Diversity and community in the twenty-first century, The 2006 Johan Skytte Prize Lecture. *Scandinavian Political Studies*, *30*, 137–174.

Rathore, S. S., Lenert, L. A., Weinfurt, K. P., Tinoco, A., Taleghani, C. K., Harless, W., and Schulman, K. A. (2000). The effects of patient sex and race on medical students' ratings of quality of life. *American Journal of Medicine*, *108*, 561–566.

Reis, H. T., and Shaver, P. (1988). Intimacy as an interpersonal process. In S. Duck (Ed.), *Handbook of personal relationships* (Vol. 24, pp. 367–389). New York: John Wiley and Sons.

Richeson, J. A., and Nussbaum, R. J. (2004). The impact of multiculturalism versus color-blindness on racial bias. *Journal of Experimental Social Psychology*, *40*, 417–423.

Richeson, J. A., and Shelton, J. N. (2007). Negotiating interracial interactions: Costs, consequences, and possibilities. *Current Directions in Psychological Science*, *16*, 316–320.

Saha, S., Komaromy, M., Koepsell, T. D., and Bindman, A. B. (1999). Patient-physician racial concordance and the perceived quality and use of health care. *Archives of Internal Medicine*, *159*, 997–1004.

Salvatore, J., and Shelton, J. N. (2007). Cognitive costs to exposure to racial prejudice. *Psychological Science*, *18*, 810–815.

Sanchez-Burks, J., Bartel, C. A., and Blount, S. (2009). Performance in intercultural interactions at work: Cross-cultural differences in response to behavioral mirroring. *Journal of Applied Psychology*, *94*(1), 216–223. doi:10.1037/a0012829

Sanchez-Burks, J., Blount, S., and Bartel, C. A. (2007). Fluidity and performance in intercultural workplace interactions: The role of behavioral mirroring and relational attunement. Working Paper No. 1039. University of Michigan.

Schulman, K. A., Berlin, J. A., Harless, W., Kerner, J. F., Sistrunk, S., Gersh, B. J., et al. (1999). The effect of race and sex on physicians' recommendations for cardiac catheterization. *New England Journal of Medicine*, *340*, 618–626.

Schwartz, L. M., Woloshin, S., and Welch, H. G. (1999). Misunderstandings about the effects of race and sex on physicians' referrals for cardiac catheterization. *New England Journal of Medicine*, *341*(4), 279–283. doi:10.1056/NEJM199907223410411

Shapiro, J., and Saltzer, E. (1981). Cross-cultural aspects of physician-patient communication patterns. *Urban Health*, *10*, 10–15.

Shelton, J. N., and Richeson, J. A. (2005). Intergroup contact and pluralistic ignorance. *Journal of Personality and Social Psychology*, *88*, 91–107.

———. (2006). Ethnic minorities' racial attitudes and contact experiences with White people. *Cultural Diversity and Ethnic Minority Psychology*, *12*, 149–164.

Shelton, J. N., Richeson, J. A., and Salvatore, J. (2005). Expecting to be the target of prejudice: Implications for interethnic interactions. *Personality and Social Psychology Bulletin*, *31*(9), 1189–1202. doi:10.1177/0146167205274894

Shook, N. J., and Fazio, R. H. (2008). Roommate relationships: A comparison of interracial and same-race living situations. *Group Processes and Intergroup Relations*, *11*, 425–437.

Sidanius, J., Henry, P. J., Pratto, F., and Levin, S. (2004). Arab attributions for the attack on America: The case of Lebanese sub-elites. *Journal of Cross-Cultural Psychology*, *34*, 403–415.

Sidanius, J., and Pratto, F. (1999). *Social dominance: An intergroup theory of social hierarchy and oppression*. New York: Cambridge University Press.

Sidanius, J., Van Laar, C., Levin, S., and Sinclair, S. (2004). Ethnic enclaves and the dynamics of social identity on the college campus: The good, the bad, and the ugly. *Journal of Personality and Social Psychology*, *87*, 96–110.

Sleath, B., Rubin, R. H., and Arrey-Wastavino, A. (2000). Physician expression of empathy and positiveness to Hispanic and non-Hispanic white patients during medical encounters. *Family Medicine*, *32*, 91–96.

Smedley, B. D., Stith, A. Y., and Nelson, A. R. (Eds.) (2003). *Unequal treatment: Confronting racial and ethnic disparities in health care*. Washington, DC: National Academies Press.

Sommers, S. R. (2006). On racial diversity and group decision making: Identifying multiple effects of racial composition on jury deliberations. *Journal of Personality and Social Psychology*, *90*, 597–612.

Son Hing, L. S., Chung-Yan, G., Hamilton, L., and Zanna, M. (2008). A two-dimensional model that employs explicit and implicit attitudes to characterize prejudice. *Journal of Personality and Social Psychology*, *94*, 971–987.

Stewart, M. A. (1995). Effective physician-patient communication and health outcomes: A review. *Canadian Medical Association Journal*, *152*, 1423–1433.

Thomas, D. A. (1993). Racial dynamics in cross-race developmental relationships. *Administrative Science Quarterly*, *38*, 169–194.

Towles-Schwen, T., and Fazio, R. H. (2003). Choosing social situations: The relation between automatically-activated racial attitudes and anticipated comfort interacting with African Americans. *Personality and Social Psychology Bulletin*, *29*, 170–182.

———. (2006). Automatically activated racial attitudes as predictors of the success of interracial roommate relationships. *Journal of Experimental Social Psychology*, *42*, 698–705.

Trail, T., Shelton, J. N., and West, T. (2009). Interracial roommate relationships: Negotiating daily interactions. *Personality and Social Psychology Bulletin, 35*(6), 671–684.

Tropp, L. R., and Pettigrew, T. F. (2005). Relationships between intergroup contact and prejudice among minority and majority status groups. *Psychological Science*, *16*(12), 951–957. doi:10.1111/j.1467-9280.2005.01643.x

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., and Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory*. Cambridge, MA: Basil Blackwell.

Van Laar, C., Levin, S., Sinclair, S., and Sidanius, J. (2005) The effect of university roommate contact on ethnic attitudes and behavior. *Journal of Experimental Social Psychology*, *41*, 329–345.

van Ryn, M., Burgess, D., Malat, J., and Griffin, J. (2006). Physicians' perception of patients' social and behavioral characteristics and race disparities in treatment recommendations for men with coronary artery disease. *American Journal of Public Health*, *96*, 351–357.

van Ryn, M., and Burke, J. (2000). The effect of patient race and socio-economic status on physicians' perceptions of patients. *Social Science and Medicine*, *50*, 813–828.

van Ryn, M., and Fu, S. S. (2003). Paved with good intentions: Do public health and human service providers contribute to racial/ethnic disparities in health? *American Journal of Public Health*, *93*, 248–255.

Welch, M. (1998). Required curricula in diversity and cross-cultural medicine: The time is now. *Journal of the American Medical Women's Association*, *53*, 121–123.

West, T. V., Pearson, A. R., Dovidio, J. F., Shelton, J. N., and Trail, T. E. (2009). *The power of one: Strength of common identity predicts intergroup roommate friendship*. Manuscript submitted for publication.

West, T. V., Shelton, J. N., and Trail, T. E. (2009). Relational anxiety in interracial interactions. *Psychological Science*, *20*(3), 289 –292. doi:10.1111/j.1467-9280.2009.02289.x

Williams, K. Y., and O'Reilly, C. A. (1998). Demography

and diversity in organizations: A review of 40 years of research. In B. Staw and R. Sutton (Eds.) *Research in Organizational Behavior*, *20*, 77–140. Greenwich, CT: JAI Press.

Wilner, D. M., Walkley, R. P., and Cook, S. W. (1955). *Human relations in interracial housing.* Minneapolis, MN: University of Minnesota Press.

Wolsko, C., Park, B., and Judd, C. M. (2006). Considering the tower of Babel: Correlates of assimilation and multiculturalism among ethnic minority and majority groups in the United States. *Social Justice Research*, *19*, 277–306.

Word, C. O., Zanna, M. P., and Cooper, J. (1974). The nonverbal mediation of self-fulfilling prophecies in interracial interaction. *Journal of Experimental Social Psychology*, *10*, 109–120.

CHAPTER 3

Policy Implications of Unexamined Discrimination

Gender Bias in Employment as a Case Study

SUSAN T. FISKE

LINDA H. KRIEGER

More than forty years after Congress passed Title VII of the Civil Rights Act of 1964, economists and legal scholars still debate whether this statute and similar others are effective and efficient tools for reducing discrimination in U.S. labor markets. Informed by principles and perspectives from neoclassical economics, some argue that all such regulation is inefficient, even counterproductive, and that markets and marketlike instruments can more effectively eliminate discrimination (Cooter, 1994; Epstein, 1995; Posner, 1987, 1989). Opposing this view, other scholars argue that market forces alone cannot eliminate all forms of discrimination from the labor market, and that at least some regulatory interventions are essential to the task (Donohue, 1986, 1987, 1989; Sunstein, 1991).

Using new behavioral and neuroscience research and using sex discrimination as a case study, we build on work challenging the rational-actor assumptions underlying much of the debate over sex discrimination law and policy (Charny and Gulati, 1998; Kang, 2005; Krieger, 1995, 1998; Krieger and Fiske, 2006). Employment decision makers (interchangeable herein with *managers*) cannot always act rationally, because, even if they consciously support equal opportunity norms, the subtle, unexamined forms of gender bias may prevent them from accurately perceiving decision-relevant information or optimally using it to make employment decisions. That is, managers might explicitly endorse equal opportunity, but unexamined prejudices might nevertheless derail their decisions. Moreover, when making employment decisions, managers' incentives often go beyond maximizing conventional utilities that, theoretically, might operate to squeeze discrimination out of employment decision making (McAdams, 1995).

Modern, Subtle Bias

In *The Declining Significance of Gender?*, Blau, Brinton, and Grusky (2006) contrast an optimistic and a pessimistic scenario for gender discrimination. On the optimistic side, they note that the progress already made will lead to more progress, that egalitarian values spread and grow, that antidiscrimination legislation is effective, that organizations are more and more female-friendly, and that women are overrepresented in the growth sectors of the economy. Indeed, women have made tremendous strides in the past century. Focusing on the world of work, our topic here, in 2006 women constituted 46% of the U.S. workforce (U.S. Bureau of Labor Statistics, 2006), and the ratio of women to men in the $25–$35,000 range was about 50-50 (Sailer, Yau, and Rehula, 2002). Upper-class women have increasing access to education, income, prestigious occupations, more egalitarian marriages, and childcare (Massey, 2007). Norms have changed dramatically, and with them, expectations about men and women have changed. The more women work full-time, the more women are seen as agentic (Diekman and Eagly, 2000), because gender stereotypes reflect the distribution of men and women into social roles (Eagly and Steffen, 1984; Hoffman and Hurst, 1990). In this view, as times change, so too will stereotypes and, consequently, discrimination.

The pessimistic view however is at least equally plausible (Blau, Brinton, and Grusky, 2006; Rudman and Glick, 2008). Popular media recount anecdotes about women opting out of paid employment (Belkin, 2003, 2005; Faludi, 1991; Story, 2005). Other pessimistic accounts acknowledge women's shift into male-dominated jobs, but also the lack of reverse

shift, whereby men stream into female-dominated jobs, suggesting that equality is stalled. Women are so occupationally segregated that 50% of men and women would have to change jobs in order for them to be equally distributed by gender (Massey, 2007). The unequal distribution of men and women across occupations accounts for about three-quarters of the earnings differential by gender. Moreover, although women increasingly work outside the home, their hours are not fully offset by men's involvement in domestic duties. Finally, formal commitments to equal opportunity do not guarantee that men and women will pursue these opportunities equally.

Indeed, the facts bear out these pessimistic accounts. Although men and women are employed to the same degree, women college graduates start out at lower salaries than men (80%) in their first year out of college (American Association of University Women [AAUW], 2007), before work histories could diverge. The more they move up the career ladder, women do even less well than men. Ten years out of college, women graduates earn 69% of male salaries. Even controlling for hours, occupation, parenthood, and other relevant factors, a quarter of the gap remains. Overall, women average 80 cents per men's dollar (Leonhardt, 2006), but in the top jobs, women's salary gap widens. Among 30,000 managers, salary raises and promotions favor men more as prestige increases (Lyness and Judiesch, 1999). For salaries in the million-dollar-plus range, the ratio of men to women is 13:1 (Sailer, Yao, and Rehula, 2002). In the Fortune 500 companies, a mere 1% of CEOs and 5% of top officers are women (Catalyst, 2002). The earnings gap between men and women continues to widen in 70% of industries sampled (U. S. General Accounting Office, 2001). Women are clearly underrepresented in managerial and high-status professions (Reskin and Ross, 1992) and earn less even when they do reach those levels (Babcock and Laschever, 2003; Reskin and Padavic, 2002; Roose and Gatta, 1999). As status increases, the male-female earnings gap increases (Massey, 2007). We do not know how much of gender disparity results from gender bias because these are correlational data, although the research does statistically control for relevant variables. Using this method, the AAUW (2007) estimates that a quarter of the pay gap is probably due to discrimination.

Moreover, causal evidence clearly demonstrates that gender bias occurs in experimental settings that control for extraneous variables. For example, in a technique known as the Goldberg (1968) paradigm, men and women present the same essay, the same résumé, or follow the same interview script, but male applicants still are favored over female ones (Eagly and Karau, 2002). A review of employment audit studies (field experiments with equivalent testers applying for advertised jobs) reached conclusions similar to the demographic correlational results (Riach and Rich, 2002). Senior, high-status positions always discriminated against women, as did male-dominated fields. Sex-integrated fields sometimes discriminated against women. Finally, although female-dominated fields discriminated against men, these jobs earned less money and carried less prestige.

Self-reports concur. Of working adults, 22% of white women, but only 3% of white men, reported discrimination (Gallup Organization, 2005). Gender was the most frequently reported form of employment discrimination (26% of the cases interviewed), surpassing even race and ethnicity. Promotion (33%) and pay (29%) topped the list. In short, gender discrimination is not over yet.

How can we account for these continuing inequalities between men and women? The answer may lie in the subtlety of the continuing forms of bias. Both the optimists and the pessimists have an argument here. Although matters have improved for women, newly recognized forms of subtle bias have been uncovered, and these help to account for the stubborn persistence of gender biases. Norms have changed about expressing prejudices, certainly for race, and to a lesser degree for gender. For example, scales of sexism created nearly four decades ago (Spence and Helmreich, 1972) are practically unusable now because almost no one endorses the most overtly prejudiced opinions (Glick and Fiske, 1996). Overt sexism is becoming as rare as secondhand smoke in polite company.

As norms have changed, techniques for measuring subtle forms of prejudice have become more sophisticated, allowing psychological scientists to demonstrate the perseverance of gender bias in less-examined forms. We describe how gender bias is, in turn, more automatic, ambiguous, and ambivalent than ordinarily understood. All of these processes likewise apply to other protected categories, notably race, although each involves some distinct processes that we note briefly. This section closes with a consideration of motivated control over these processes.

The subtle forms of bias matter partly because organizational hiring and promotion relies on managers' subjective judgments, which more easily fall prey to stereotypic assumptions (Heilman and Haynes, 2008). Even when organizations use formal criteria, we will see that stereotypes can infect seemingly objective standards. Unexamined biases all too easily taint workplace judgments, absent the systemic interventions described later.

Automatic Stereotyping

Managers cannot be blind to gender, which shapes responses from the first moments of any encounter. According to neural evidence, people identify the gender of another person in a fraction of a second (e.g., Ito and Urland, 2003). Rapid gender-related associations follow in another fraction of a second. These automatic, often unconscious associations carry cognitive beliefs (including stereotypes), affective reactions (including emotional prejudices), and behavioral tendencies (including discrimination). Some consequences of the relatively automatic biases include, as we will see, within-gender category confusions, the priming of accessible gender biases, implicit gendered associations, and gender-category application under cognitive load (see Fiske and Taylor, 2008; Fazio and Olson, 2003, for reviews).

CATEGORY CONFUSIONS

After people rapidly identify each other's gender, they use this category to sort people and to tag the ensuing information. As a result, people tend to confuse other people who fall into the same category, forgetting *which* woman (or *which* black or *which* older person) contributed a suggestion. In the who-said-what experiments, spontaneous memory errors more often confuse people within a category than between categories (Taylor et al., 1978). At least twenty studies have demonstrated this confusion within-gender categories, as well as with race, age, sexual orientation, attitudes, attractiveness, skin tone, and relationship type (See, e.g., Maddox and Chase, 2004; Maddox and Gray, 2002; for reviews, see Fiske, 1998, pp. 371–372; Klauer and Wegener, 1998). These memory errors occur without apparent intention, effort, or control, making them relatively automatic.

These confusions are not harmless; they encourage gender-based stereotyping (Taylor et al., 1978). That is, when behavior is tagged by gender, its interpretations shift: tagged as female, warm behavior is motherly; tagged as male, it is socially skilled. Aggressive behavior, tagged as male, might be assertive, but tagged as female might be bitchy. These stereotypic interpretations and category-constrained associations operate below the surface. The category confusion itself may be all-too-painfully conscious, but it is subtle, first because people are typically unaware that they are lumping all the women into one category and all the men into another.

Second, the subtle harm of category-based confusion operates particularly in a male-dominated workplace. If managers fail to distinguish among women, it sends a message that they are interchangeable and not individuals worth differentiating. If the men seem interchangeable, it merely conveys that they are the contextual default group. As the proportion of one gender increases, the group as a whole is stereotyped according to that gender (Taylor et al., 1978), in this case masculine ones. Given the higher status of the average male-dominated workplace compared with a female-dominated one, confusing the women with each other undermines their being identified as individual prospects for better jobs, but in a manner that is difficult to identify, since it is subtle.

Third, people are unaware of the category's effects on their interpretations. Category confusions and the resultant stereotypes penalize women at work because the default female is stereotyped as nice but incompetent, therefore especially unsuited to high-status employment. And the professional female subtype is stereotypically competent but cold (Glick et al., 1997), not conducive to promotion, as the first author testified in a relevant discrimination case (Fiske et al., 1991; *Hopkins v. Price Waterhouse*, 1985). Even if category confusions cause men to be stereotyped as masculine, male stereotypes are a closer fit to high-status jobs, such as manager (Heilman, 1983), so gender stereotypes harm women's advancement more than that of comparable men. Category confusions, and the underlying process of categorical thinking, reflect subtle but harmful automatic bias at work.

ACCESSIBILITY

Another form of automatic bias occurs when people encounter a *prime* that makes accessible associated material in memory. Whether the prime itself enters one's mind consciously or unconsciously, this priming process occurs without awareness, so it illustrates another route for gender bias to operate under the radar in settings that prime stereotypic roles. For example, gendered language can prime stereotypic interpretations of women's traits, to their detriment in the workplace; as noted, stereotypically feminine attributes are deemed liabilities for jobs with prospects for advancement.

Priming effects occur through the accessibility of certain gender-occupation associations (McConnell and Fazio, 1996). That is, the frequent suffix *-man* added to congressman or businessman influences the inferences drawn about the most appropriate gender for that job. The allegedly neutral *-person* suffix actually enhances the accessibility of female stereotypes. Enhanced accessibility of masculine and feminine attributes occurs partly through frequent exposure to occupational-title suffixes (consider *governor* versus *governess*). The accessibility of primed attributes influences perceptions of the target person whose gender

makes the prime applicable, again to the detriment of women in high-status work.

One of the earliest accessibility studies demonstrated how cognitively primitive this process can be. When people categorized another person as one of "us," they automatically recognized positive traits more rapidly than when they categorized another person as one of "them" (Perdue et al., 1990). People engaged in a lexical decision task (judging letter strings rapidly as words or nonwords) identified positive words preceded by *us* faster than those preceded by *them*. This so-called priming of compatible responses (we = good; they = less good) appear in numerous studies of automatic associations (see Fiske and Taylor, 2008). To the extent that a workplace is male-dominated, an entering minority of women likely elicit responses that "they" are "less good" than "we" are.

Overall, priming studies have demonstrated that words related to women prime associations that reflect gender prejudice. For example, priming with gender-associated words (doctor, nurse) speeds identification of gendered pronouns, regardless of people's awareness of the prime-target relations and regardless of their explicit beliefs about gender (Banaji and Hardin, 1996). Similar effects link pictures of men and women, or even masculine and feminine objects (e.g., baseball mitt versus oven mitt), to gender-stereotypic occupations (Lemm, Dabady, and Banaji, 2005). Priming thus affects not only the speed of identifying a relevant stimulus as a male or female person; it also affects stereotypic interpretations. That is, priming with stereotypically female behaviors (e.g., "won't go alone") made participants rate women as more dependent than men who performed the same behaviors (Banaji, Hardin, and Rothman, 1993). The typical contextual prime is not equal opportunity but sticks to stereotypically applicable stimuli.

Illustrating a different dimension, men primed with sexually suggestive advertisements subsequently, on a lexical decision task, responded faster to sexist words and more slowly to nonsexist words compared with controls (Rudman and Borgida, 1995). In a subsequent job interview of a female candidate, sexually primed men asked her more sexist questions, recalled better her physical appearance, wrote less about her job qualifications, and rated her as less competent but more friendly. For those men who also scored high on an individual difference measure of likelihood to sexually harass, the primed ones were actually more likely to recommend hiring her as an office manager; however, their reasons for hiring her may have been reflected in the confederate interviewees' own ratings of the primed interviewer. Despite being blind to condition and individual differences, the women more often judged primed interviewers as staring at their bodies, as sexist, and as sexually motivated. Independent judges (also blind to experimental conditions) likewise rated primed men as more sexual, more dominant, and as sitting too close. The interview context makes this paradigm especially relevant to employment selection in male-dominated workplaces that display sexually provocative materials.

THE IMPLICIT ASSOCIATION TEST

Priming focuses on a sequence whereby a prime makes accessible a subsequent interpretation of a relevant stimulus. Many measures of cognitive association tap frequently paired concepts occurring simultaneously without assuming that one comes first. The Implicit Association Test (IAT) is widely used to measure a variety of socially or politically sensitive associations (Greenwald et al., 2002; Greenwald, McGhee, and Schwartz, 1998).

Because the IAT is famous and focuses on prejudices, it has provoked controversy. Research using the IAT shows considerable convergent validity (i.e., relating it to other subtle measures; Cunningham, Preacher, and Banaji, 2001); the IAT also shows predictive validity (e.g., relating it to behavior; Greenwald et al., 2009). Most people show essentially automatic favoritism on the basis of common social categories such as gender, race, age, religion, nationality (Rudman et al., 1999), and even minimal, arbitrary group memberships (Ashburn-Nardo, Voils, and Monteith, 2001).The IAT can distinguish between liking and respect. Beside sheer liking, one study examined attitudes toward female authority using the IAT, priming measures, and more explicit attitude measures (Rudman and Kilianski, 2000). The IAT asked people to categorize a series of words according to either a congruent rule (high-authority words related to men versus low-authority related to women) or an incongruent rule (the opposite). Participants saw individual words and classified them by one of two computer keys. Faster responses to congruent pairings indicate closer mental associations than the slower responses to incongruent pairings. In this study, implicit attitudes toward female authority and gender beliefs, both measured by the IAT, correlated with each other and with priming measures. Explicit attitudes also correlated with each other, illustrating the simultaneous existence of consciously egalitarian ideologies and less conscious automatic stereotypic associations in memory.

For other groups, too, implicit attitude measures correlate with each other and with nonverbal behavior, all being unexamined responses. Because prejudice is a sensitive topic, implicit attitudes correlate

only sometimes with explicit attitude measures, which correlate with each other and with verbal (controllable) behavior (Dovidio, Kawakami, and Gaertner, 2002; Greenwald et al., 2009; Hofmann et al., 2005). In a related technique, indirect priming measures also predict nonverbal behavior in interracial interactions, evaluations of an out-group member's essay, emotional reactions to out-group members, and other relevant attitudes (Fazio and Olson, 2003).

Implicit measures may reveal associations that people would rather not admit. As a gender example, agentic (authoritative) women elicited a backlash for violating prescriptive stereotypes that women should be nice (Rudman and Glick, 1999). This prescription reflects people's agency versus communality stereotype, but only on the implicit, not the explicit, level (Rudman and Glick, 2001). These implicit associations are real in their consequences as an interpersonal skills issue, for example, in demanding that a high-powered female accounting executive behave in a more feminine manner (Fiske et al., 1991; *Hopkins v. Price Waterhouse*, 1985). On a different dimension, but one that is also work related, people implicitly associate men with science and women with liberal arts and family (Nosek, Banaji, and Greenwald, 2002b).

The IAT reflects an automatic process, but only relatively so, in that people are aware of the words and pictures they are pairing, and they may even become aware of their relative speed of associating stereotypic combinations faster than counterstereotypic combinations. Indeed, this makes the IAT a useful didactic device. But people are not aware of the processes that speed or slow their responses. The IAT is only relatively automatic, too, in that it varies with context. That is, perceivers may focus more on a person's race or gender, cuing different associations (Mitchell, Nosek, and Banaji, 2003). Or, with effort, perceivers may form counterstereotypic images (Blair, Ma, and Lenton, 2001). Perceivers can control these effects if they have a counterstereotypic intention and if they have enough time and motivation (Blair and Banaji, 1996), although sometimes this effort will backfire (Frantz et al., 2004).

In general, as mentioned, the IAT correlates moderately with explicit measures, according to meta-analysis across studies (Hofmann et al., 2005). Correlations increase when explicit self-reports are more spontaneous and when the measures are conceptually most similar. As noted, implicit measures correlate with nonverbal behavior, which is less monitored but still can create a chilling effect. Overall, the IAT may have particular utility for attitudes that people hesitate to report explicitly. And gender prejudice is one such attitude.

The generality of implicit measures to other sensitive topics is clear. Just as gender shows automatic biases, so does race. Indeed, racial bias studies were among the first to identify truly automatic forms of bias by methods that predate the IAT. Whites identify positive traits (e.g., "smart") faster when first primed with *whites* than with *blacks* (Dovidio, Evans and Tyler, 1986; Gaertner and McLaughlin, 1983; Perdue et al., 1990). People's automatically positive in-group associations recur reliably over time (Kawakami and Dovidio, 2001) and predict nonverbal behavior in interracial interactions (Dovidio, Kawakami, and Gaertner, 2002). (Overt attitudes predict overt verbal behavior.) People can control their stereotypic associations by extensive training that is specific to the particular out-group, but it does not generalize to other out-groups (Kawakami et al., 2000). Note that many of these effects emphasize the in-group positives more than derogating the out-group negatives.

Across groups then, the IAT shows one more way that people have biases that they cannot easily acknowledge. Subtle biases matter in job-interview experiments, for example, where unexamined race-related nonverbal behavior creates social distance and damages interviewee performance and contributes to the interviewees' perception of the interviewer as unwelcoming (Word, Zanna, and Cooper, 1974). Along with category confusions, primed accessibility, and implicit associations comes one last type of relatively automatic bias.

CATEGORY ACTIVATION

People notice a person's gender faster than they can say the person's name (Cloutier, Mason, and Macrae, 2005; Macrae et al., 2005). Although it is a relatively automatic process, category activation precedes the attention to cues relevant to multiple alternative categories (e.g., both gender and age), but people activate only the currently most relevant category (Quinn and Macrae, 2005) or the most accessible one (Castelli et al., 2004).

Having activated a category, people most easily process stereotype-consistent meanings, which take less cognitive capacity to process. But stereotype-inconsistent information especially bothers people who are prejudiced; they especially attend to stereotype-inconsistency in an effort to explain it away (Sherman, Conrey, and Groom, 2004; Sherman et al., 1998, 2005). At early stages of stereotype application, people prioritize a coherent impression, so they work on the inconsistencies; they then remember the inconsistent information that had required cognitive work to assimilate. However, because they have

explained or assimilated it somehow, the inconsistency may well not undermine their stereotypes.

SUMMARY OF AUTOMATIC STEREOTYPING

The first leg of subtle stereotyping, its automaticity, rests on basic categorization and rapid association processes. People easily tag other people by gender, confusing women with other women or men with other men and being biased accordingly. Because workplace status correlates with gender, this most often disadvantages the women. People's unintended biased associations emerge in stereotype-relevant cognitive tasks, showing the facilitation of stereotypic associations and prejudices. These relatively automatic processes do vary with cognitive load and motivation, mostly in the service of efficiency and stereotype maintenance. Stereotypes of women operate to their disadvantage in the workplace. Automatic racial and other biases operate in similar ways. Whether stereotypes and evaluations measured this way are the "real" attitude or not matters less than the observation that these upstream, relatively automatic indicators do correlate with downstream consequential attitudes and behaviors.

In the workplace, if dominant groups more often hold powerful positions, their automatic biases matter more than the automatic biases of their subordinates. What is more, people in power are especially prone to automatic stereotyping for several reasons (Fiske, 2010b). Power disinhibits behavior, making people monitor themselves less and operate on automatic more. Powerholders need each individual subordinate less than vice versa, so their motivation to be accurate is lower. Stereotyping other people is easier than individuating them, and for powerholders, it has fewer readily recognizable negative consequences.

Ambiguous Stereotyping

Subtle stereotyping is difficult to detect not only because it is fast, but also because it is slippery. Interpretation is everything. People deal automatically with clearly stereotypic or counterstereotypic information, as just seen. People also interpret ambiguous information to confirm their biases. For example, subjects were primed subliminally (i.e., below awareness) with rape-related terms (e.g., *rape*, *aggressive*, *scream*), and then read an ambiguous aggressive sexual encounter between a man and a woman, one in which the responsibility for the sexual interaction was mixed (Murray, Spadafore, and McIntosh, 2005). Rape-primed observers afterward judged the woman more negatively than did unprimed observers, but only if they believed generally that the world is a just place (i.e., that people get what they deserve). Observers spontaneously attribute responsibility for events, and the rape-related primes presumably made the just-world believers more likely to blame the female participant for the ambiguous encounter. Those low on belief in a just world had the opposite reaction, rating her more positively, presumably because they did not blame her. Although the specifics might apply most closely to workplace interpretations of perceived responsibility for sexual harassment, the larger point of the example is that most behavior is ambiguous and open to interpretations depending both on observers' belief systems and on contextual primes. These effects on interpretation of ambiguous behavior are subtle and unexamined.

People interpret not only the content of ambiguous information but also its causal meaning. This phenomenon has obvious applications to employment evaluations. That is, when men succeed at traditionally masculine tasks, this success is viewed as reflecting their inherent ability or worth, but when women achieve at the same (male) task, success is attributed to chance or circumstance (Deaux and Emswiller, 1974; Swim and Sanna, 1996). The reverse obtains for traditionally feminine tasks. Applied to in-group–out-group relations generally, this phenomenon is termed the ultimate attribution error (Pettigrew, 1979), and the effect appears in interethnic attributions (Hewstone, 1990). Recent evidence qualified this pattern, showing that it applies only to stereotype-confirming attributions (Glick et al., in preparation).

Interpretations of ambiguous information do not always disadvantage women in the most obvious ways. Because of "shifting standards," a woman may receive praise when she performs well (for a woman). A man would have to perform better (well, for a man) to receive equal praise, because the male standards are higher for traditionally male domains. However, one cannot live on praise alone. When allocating scarce resources, that is, when men and women compete, then the same stereotypes result in pro-male bias (Biernat and Vescio, 2002). Shifting standards could favor women on qualitative judgments in a masculine task (rating them as "smart") but could disfavor women on a quantitative judgment about the same domain (ranking the smartest people in the organization). Shifting standards appear in contexts from sports (Biernat and Vescio, 2002) to military promotions (Biernat et al., 1998).

These tacit processes capitalize on the ambiguity of the information given, so the influence of the bias itself is subtle and difficult to detect. People hide their biased interpretations from themselves as well

as others. For example, in evaluating job applicants, people weight the applicants' credentials differently depending on how they want the decision to come out (Norton, Vandello, and Darley, 2004; see also Uhlmann and Cohen, 2005). Men evaluating men and women for a construction job weighted education more for men, and overall, unless the woman was more educated, in which case they shifted to experience as the more important criterion. What they valued in one decision they then carried over to subsequent decisions. In a similar vein, people who have proven their moral credentials as unprejudiced (rejected sexist statements) subsequently feel more free to express stereotypes and prejudices (hire the man for the construction job) (Monin and Miller, 2001).

SUMMARY OF AMBIGUOUS STEREOTYPING

The ambiguity of many category-based cognitive biases is clear. People interpret ambiguous information to fit their biases, both unconscious and conscious. People's causal attributions for men's and women's behavior reinforce their biases. Because of lower standards for women on masculine tasks ("good for a woman"), qualitative judgments may actually favor women, for example, when allocating praise or shortlist spots, but allocation of scarce resources (e.g., ranking or hiring) does not. People justify their discriminatory judgments to themselves and to others, making their biases ambiguous.

Ambivalent Biases

Subtle biases also hide behind ambivalence. Unlike automaticity, which becomes evident as more explicit, controlled biases recede, and unlike ambiguity, which people use to excuse or disguise their own biases that are aversive to themselves and to others, ambivalence is not new. Biases often veer between disliking and disrespecting. Women who inhabit traditional subordinate roles (housewife, secretary) are often liked but disrespected, whereas those who inhabit nontraditional roles are often respected but disliked (career women, feminists) (Eckes, 2002; Glick et al., 1997). Ambivalent sexism (Glick and Fiske, 1996, 2001) identifies these two factors: first, hostile sexism, namely resentment directed toward women who pursue nontraditional roles, gaining respect but forfeiting affection, and second, subjectively benevolent sexism directed toward women who stay within prescribed gender roles, gaining protection but foregoing respect. This time-honored pattern holds up across cultures (Glick et al., 2000).

On an individual level, ambivalent sexism relates to a variety of discriminatory judgments about the hiring of women (Masser and Abrams, 2004; Uhlmann and Cohen, 2005) and sexual harassment on the job (O'Connor et al., 2004; Russell and Trigg, 2004). Benevolent sexism predicts favorable evaluations of women in traditional roles (homemaker), but hostile sexism predicts unfavorable evaluations of women in nontraditional roles (career women) (Glick et al., 1997).

The generality of ambivalence as a foundation of intergroup perception suggests that mixed reactions inhere in most biases. As one example, racial ambivalence (Katz and Hass, 1988; Katz, Wackenhut, and Hass, 1986) reflects White liberals as mixing interracial reactions pro and con, viewing Black people alternatively as disadvantaged because of external obstacles, such as discrimination, or internal obstacles, such as values and motivation.

Summary of Subtle Gender Biases and Comment on Other Group Biases

As recent research shows, subtle forms of bias are more automatic, ambiguous, and ambivalent than laypeople expect. Gender has been our case study here, but race shows essentially parallel effects, although with some distinct patterns. For example, people's own potential for racism is more aversive to them than is their potential sexism (Czopp and Monteith, 2003), so if anything, the automaticity of racism masks more internal conflict than sexism does. Expressing gender bias is less problematic than expressing race bias, in part because gender distinctively combines men's societal dominance with men's and women's intimate interdependence (Glick and Fiske, 2001). Race is distinct in its history, entailing forced immigration and slavery, with continuing consequences in neighborhood segregation and social-class disparities (Fiske and Taylor, 2008, chap. 12). Other biases based on age, disability, or religion, for example, each have unique features, but most principles of unexamined biases—being automatic, ambiguous, and ambivalent—probably occur in most cases.

The Challenge of Controlling Subtle Forms of Bias

Aside from the well-known blatant types, modern forms of gender bias emerge as automatic, ambiguous, and ambivalent. The relatively unconscious, unintentional, murky, and mixed signals are difficult for perceivers to notice and for targets to interpret. Nevertheless, they have a chilling impact on employment settings, as already indicated. What are the

prospects for controlling subtle gender biases, or at least for controlling the discriminatory decisions that flow from them? And what interventions—including intra-organizational reforms, regulatory initiatives, private or public enforcement actions, or the use of markets and marketlike mechanisms—are best, or at least well, suited to achieving this policy objective?

The Trouble with Subtle Bias: Failures of Motivated Control

If biases are so often subtle, how can people avert them? The greatest obstacle to motivated control would seem to be the automaticity of bias. However, if biases qualify as only relatively automatic, then with sufficient motivation, capacity, practice, and information, people should be able to avoid them (Blair and Banaji, 1996; Kawakami et al., 2000; Kawakami, Dovidio, and van Kamp, 2005; Macrae et al., 1997).

In fact, a variety of motivational strategies do appear to undermine automatic biases. These include perspective taking (Galinsky and Moskowitz, 2000), guilt (Son Hing, Li, and Zanna, 2002), self-focus (Dijksterhuis and van Knippenberg, 2000; Macrae, Bodenhausen, and Milne, 1998), and affiliative motivation (Sinclair et al., 2005). Motivation and capacity cooperate to encourage individuation (for a review, see Fiske, Lin, and Neuberg, 1999). Motivated control can in extraordinary circumstances override even automatic stereotypes. Presumably, managers could be motivated by perspective taking, guilt, and so on, to make more individuated decisions.

Reliance on self-control presupposes that people notice their bias. However, decision makers cannot always know when gender (or any other protected category) has tainted their responses. Arguably, most perceivers would be motivated not to notice their own biases. To view oneself as biased endangers one's self-image as an objective, fair-minded person. Most people think they are less biased than other people (Pronin, Lin, and Ross, 2002), but most people cannot be above average. All the evidence suggests that people are ignorant of their everyday biases. Specifically, people do not so much fail to control their biases as fail to notice that they have them (Amodio, 2008). Thus, we cannot count on perceivers to monitor themselves successfully. It takes structured organizational processes to facilitate their doing so, as we will see.

More Trouble with Subtle Bias: Targets' Inability to Recognize and Their Reluctance to Claim Discrimination

Ironically, even targets of bias may not be the best judges of when they are subjected to discrimination. First, people are not their own control group. People rarely get to see how they would be treated if they were the other gender: in legalese, "but for" being female. One is not usually able to observe fully the treatment of one's "male comparator," the similarly situated opposite-gender other. Anyway, people always differ in more ways than gender, so one's individually matched control person is hard to find. In short, targets cannot usually know when they have been discriminated against, partly because of the difficulty inherent in disentangling the reasons—discriminatory or nondiscriminatory—why a particular decision was made.

This dilemma is exacerbated when patterns of disparate treatment are inconsistent or inconspicuous and therefore complicated. Also challenging are occasions that do not violate principles of *ordinal equity* (Rutte et al., 1994). Ordinal equity is violated when a person who ranks higher on relevant input variables (e.g., qualifications) ranks lower than another person on the relevant outcome variable (e.g., salary, rank, or grade). *Magnitudes* of difference in the input or output variables do not implicate ordinal equity. Thus, a person may receive inadequate compensation that does not reward merit proportionately to others on the same scale.

People also have difficulty identifying discrimination when it manifests as small, seemingly insignificant forms of preference or leniency toward members of one particular group. When laypeople think of discrimination, they consider discrimination *against* members of a group who are treated unfavorably according to some standard metric of fairness. But discrimination can also occur when members of the subordinated group are treated neutrally according to some standard metric of fairness, while members of the privileged group are treated with greater leniency or favor than neutral rules would direct (Brewer, 1996). When these subtle forms of advantage accrete over time, they can be difficult to detect (Krieger 1998).

Even when discrepancies are evident, admitting that one has been the victim of discrimination is undesirable. People do not as readily "play the gender card" as critics imagine. Targets are reluctant to attribute negative outcomes to prejudice because of the social as well as personal costs. Socially, people who make attributions to discrimination risk being labeled as complainers, troublemakers, or worse (Kaiser and Miller, 2001, 2003; Swim and Hyers, 1999), especially by highly prejudiced people (Czopp and Monteith, 2003). An interaction is irrevocably altered when one person raises issues of prejudice. People want to belong to their group (whether work or social), and allegations of discrimination undermine

belonging; hence, people high on a need to belong are less likely to make attributions to discrimination (Carvallo and Pelham, 2006). Attributions to discrimination also can undermine the target's personal sense of control. Although attributing negative outcomes to discrimination can buffer self-esteem, attributing positive outcomes to (bend-over-backward positive) discrimination can undermine self-esteem because then one cannot take credit for the success (Crocker et al., 1991). Constant attribution to discrimination damages self-esteem (Major, Testa, and Blysma, 1991). And perceiving one's personal experience as an instance of pervasive discrimination is depressing (Schmitt, Branscombe, and Postmes, 2003).

Of course, perceived discrimination sometimes reflects reality, and this is the target's predicament. People and situations vary in the extent to which they resolve attributional ambiguity by playing the discrimination card. Low-status group members (including women) who endorse individual effort and the Protestant ethic are *less* likely to make attributions to discrimination (Major et al., 2002). On the other hand, high-status group members (men) who hold the same beliefs are *more* likely to attribute their negative outcomes to discrimination against themselves. Women may be less likely than men to make attributions to discrimination perhaps because their status is less legitimated, so frustration and disappointment are familiar. Attributions to discrimination interact with expectations about entitlement and desert. In summary, targets face a number of obstacles to identifying discrimination, including the lack of a personal control group, the complexity of individual evidence, in-group leniency, causal ambiguity, and the negative effects of attributions to gender bias.

The Inefficiency and Inefficacy of Individual Disparate Treatment Adjudication

As the second author of this chapter has extensively described in earlier work (Krieger 1995, 1998), the individual disparate treatment discrimination case, in which a single plaintiff accuses an employer of intentional discrimination, is an extremely weak policy tool for addressing subtle forms of discrimination in the labor market. As just described, discrimination resists identification from single-case data. In any individual sex discrimination case, many plausible reasons other than the plaintiff's gender may account for the employer's action. Teasing apart the causal roles of nondiscriminatory and discriminatory motivations is in most cases difficult, expensive, and risks both Type I (as noted, overidentifying) and Type II (underidentifying) errors (Krieger, 1995, 1998; Wax, 1999).

Why would one expect a trier of fact to approach this attribution task free of the same forms of subtle bias that the suit seeks to expose and remedy? Such a conclusion would be reasonable only if one believed that levels of implicit bias among fact finders differed from those among employment decision makers. No apparent evidence justifies this position. Consequently, we can assume only that the outcomes for individual disparate treatment cases will, at least in significant part, fall victim to the very same biases that generated the very grievances being adjudicated.

Despite the inefficiency and inefficacy of individual disparate treatment litigation in equal employment opportunity (EEO) policy enforcement, it is, at present, the primary policy tool used to accomplish the task. While class action litigation and affirmative action initiatives (both voluntary and mandatory) dominated the EEO policy landscape during the 1970s, their influence waned suddenly and dramatically during the Reagan and first Bush administrations (Kalev and Dobbin, 2006; Krieger, 2007). By 1991, when the economists John Donohue and Peter Siegelman reported changes in the nature of federal employment discrimination litigation, class action hiring and promotion discrimination lawsuits had almost completely disappeared from the legal landscape. They had been almost completely supplanted by individual disparate treatment cases (Donohue and Siegelman, 1991).

Summary of the Trouble with Individual Solutions

Subtle biases defy individualized solutions for three main reasons. First, decision makers typically do not notice subtle (hence, unexamined) biases. Even if they do, motivated control both creates mental overload, endangering other tasks at hand, and creates stereotype rebounds. Second, the targets themselves may not recognize the subtle biases that disadvantage them, because they do not have a personal control group for comparison. The effects of subtle bias are difficult to detect at the individual level, and the causal antecedents of negative employment outcomes are inherently ambiguous. Also, targets resist viewing themselves or being viewed as victims of discrimination because of the negative ramifications for self and others. Finally, accumulating evidence indicates that the individual disparate treatment claim, now a favored tool in antidiscrimination policy, is not particularly effective. What are we to do?

Modern Policy Tools for Modern Forms of Bias

From here, our attention turns to potentially effective policy prescriptions. If the individual disparate treatment adjudication is an inefficient and ineffective tool for reducing the levels of discrimination caused by subtle forms of gender bias, what might supplement

it?[1] We consider two sets of policy initiatives here: (1) voluntary measures by organizations to reduce discrimination caused by uncontrolled application of unexamined subtle bias; and (2) renewed and extended enforcement of EEO compliance programs (such as those administered by the 1970s Office of Federal Contract Complicance Programs [OFCCP]), supplemented by mandatory, public, EEO-compliance disclosure requirements, analogous to those now used in the regulation of financial markets. Although untested, these solutions follow from existing research and point to new research directions. They deserve to be true, though whether, when, and how they will work remains an empirical question.

Organizational Initiatives

Organizations can implement some solutions to mitigate the impact of unexamined biases by monitoring and motivating their members and by providing them with information and structured decision-making tools, according to current research.

MONITORING

Individuals are not in the best position to notice discrimination against themselves because they do not have the necessary perspective, including full comparisons to similarly situated others, as noted earlier. Big-picture perspective matters also in that exposure to multiple instances reveals patterns that individual instances do not. A survey of working men, working women, and housewives indicated that although women were objectively objects of discrimination, they recognized discrimination only in the aggregate, but not in their own particular case (Crosby, 1984). The women blamed themselves and were uncomfortable identifying themselves as targets of discrimination, even when they acknowledged discrimination against women in general. Perhaps, then, the organization is in a better position to monitor than the individual is. The necessity for organization-level monitoring becomes even clearer in a follow-up study showing that people easily perceive patterns of discrimination in the aggregate but not when information appears on a case-by-case basis (Crosby et al., 1986).

Organizations routinely gather at least some aggregate data on gender in work-force composition, promotion, and pay to report to the Equal Employment Opportunity Commission (EEOC), as required by federal EEO law. It is not a stretch to expect that their human relations departments can, should, and often do already monitor disparities. However, as the second author of this chapter has detailed in earlier work (Krieger, 2007), the scope and quality of those monitoring efforts have greatly diminished since the late 1970s, and government agencies rarely hold employers responsible for any disparities that they uncover. We will return to this subject later.

MOTIVATION

Conventional rational-actor models assume that people are motivated mainly by self-interest, but these models are giving way under the influence of behavioral and neuro-economic evidence (e.g., in this volume: Jolls; Loewenstein and John; Sunstein and Thaler; Tyler; Weber). Here, to expand the discussion of motives that matter, we will address a range of motivations relevant to undermining discrimination. These motives draw on a social evolutionary framework (Fiske, 2010a) that also fits the history of the most frequently and prominently identified motives in psychological science (Fiske, 2007).

Harnessing the Motive to Belong

People are motivated to create social ties with others; indeed, other people are the most powerful evolutionary niche. That is, people have always survived and thrived better in social networks than in isolation. The immunological and cardiovascular risks from social isolation culminate in an age-adjusted mortality threat that is equivalent to cigarette smoking (Fiske, 2010a, chap. 1). As noted, much of the variance in prejudice results from favoring one's in-group (e.g., leniency toward those of one's own gender). In-group bias inevitably disadvantages the out-group, whether based on gender, race, age, or other salient identities. People are most comfortable with others they perceive to be like themselves, so organizations spontaneously create homogeneous environments if left to themselves (Gruenfeld and Tiedens, 2010). Belonging motives therefore tend to maintain the segregated status quo.

Belonging motives nonetheless can be harnessed to overcome employment segregation and discrimination by changing the perceived boundaries of the group (Estlund, 2000). If people believe that the relevant in-group crosses gender boundaries, for example, then they are less likely to discriminate on the basis of gender. Departmental and corporate identity can foster a common in-group identity (Gaertner and Dovidio, 2005). Social categorization that crosses boundaries (i.e., profession crosscutting gender) also diminishes discrimination (Brewer, 2000).

Consider several examples of how organizations might do this:

- Make diversity part of the organization's identity and mission; a mixed-gender, multicultural client base requires a mixed-gender multicultural work force

- Emphasize pragmatic identities (e.g., teams) within the organization, crosscutting demographic categories
- Promote the organization's history of successfully diversifying its employees

Harnessing the Motive to Understand

To survive and thrive in their groups, people aim for accurate understanding as the group defines it. One cannot operate socially except within the group's defined reality. Socially shared understanding means that informational influence (perceived fact) among group members is identical to normative influence (consensus) among group members (Turner, 1991). That is, people believe their groups to have an accurate understanding of reality. Thus, what the group defines as true about itself and other groups will prevail.

Applied to gender discrimination, socially shared understanding communicates shared beliefs about gender stereotypes and prejudices, men's and women's suitability for particular jobs, the prevalence and meaning of affirmative action, the possibility of discrimination, and whether diversity serves the best interests of the organization. Local group norms about gender determine the social truth (Prentice and Miller, 2006).

What's more, people seek efficient understanding, and the simplest adequate answer will prevail; people are cognitive misers (see Fiske and Taylor, 2008, for references to these processes). Accordingly, as noted, they often use stereotypes, especially when cognitively overloaded. Although they can be motivated to go beyond their stereotypes, in the absence of motivation they do not take the trouble. Motives for accurate understanding over rapid understanding, however, can take people beyond their stereotypes (e.g., Fiske, Lin, and Neuberg, 1999).

People also use stereotypes when criteria are vague. Hence, socially shared, empirically validated, measurable criteria can undermine subjectivity. When decision makers operate on automatic, they do not notice their subtle biases, but encouraging people to consider the opposite can jolt them out of their well-travelled decisional groove (Lord, Lepper, and Preston, 1984).

Consider how organizations might harness the motive for shared understanding:

- If all norms are local, communicate that traditional gender stereotypes do not apply in this setting
- Educate decision makers that their evaluative judgments may be influenced by subtle forms of bias, no matter how sincere their conscious commitment to fairness, objectivity, and nondiscrimination
- Reduce, where possible, decision makers' levels of cognitive busyness
- Generate clear, objective evaluative criteria and supply evaluators with information relevant to those criteria
- Encourage decision makers to "consider the opposite" before acting on or attempting to justify an initial impression

Harnessing the Motive to Control

One motive for accurate understanding is control over one's own outcomes. People seek efficacy in controlling the contingencies between their behavior and their outcomes. On this point, psychologists (especially reinforcement theorists) and economists agree. If a manager's promotion depends on the ability to hire and promote underrepresented groups, the manager will find ways to do so.

Organizations can take affirmative steps to assist managers in reducing the amount of discrimination resulting from implicit bias. Much of the existing bias reduction research draws on a *dual process model* of social cognition (Chaiken and Trope, 1999). In making sense of others, people use two systems of information processing: a "low-road" system, which is automatic, rapid, unconscious, and demanding of few cognitive resources, and a "high-road" system, which is effortful, conscious, controlled, and resource intensive.

Social psychologists agree that stereotyping includes both automatic and controlled components. If a person has learned but then consciously rejects the stereotypes and attitudes associated with a devalued social group, those stereotypes and attitudes do not just disappear. Indeed, they can be measured by techniques such as the Implicit Association Test, even in many people who rate low on measures of explicit bias or who are themselves members of the negatively stereotyped group (Dasgupta 2004; Nosek, Banaji, and Greenwald, 2002a). In such situations, the older implicit attitudes and associations continue to exist alongside the newer, consciously held beliefs and commitments. The implicit attitudes function as a low-road system, while the conscious beliefs function as a high-road system. In fact, the neurological substrates underlying the implicit-explicit distinction appear through such technologies as functional magnetic resonance imaging (fMRI) (Lieberman et al., 2002).

As described earlier, relative consensus within cognitive social psychology holds that stereotype *activation*, when it occurs, is automatic. But a growing consensus also holds that stereotype activation does not *necessarily* lead to stereotype expression and that stereotype activation itself can be affected by environmental factors. Even where a social expectation (such

as a stereotype) is activated and generates a schema-consistent impression, the application of conscious, high-road thinking can override that initial impression (Monteith and Voils, 2001). However, certain conditions must be present for this to occur (Bargh, 1999; Wilson and Brekke, 1994). Perceivers must be *aware* of the possibility that their initial impression might be biased, and they must be *motivated* to correct for that bias. For correction to occur, perceivers must also have time and attention to spare: like all controlled processing, high-road correction requires cognitive resources. And finally, to correct an initially biased impression, perceivers must have available the information and analytical tools required for meaningful deliberation.

Consider how the presence or absence of these factors might influence the degree of control over the gender bias present in hiring decisions. If those who make hiring decisions are simply told to be "gender-blind" in evaluating potential candidates, if they do not believe themselves to be gender biased and if there are no meaningfully endorsed and enforced system of goals relating to hiring and promoting women, conscious correction of implicitly gender-biased evaluations is unlikely to occur. Under such conditions, managers are unlikely to be aware of the possibility that implicit gender stereotypes and attitudes could be skewing their judgment, and they will have little motivation to engage in conscious correction.

In the hiring or promotion context, at least with respect to high-level jobs, the ultimate decision making is often quite deliberate, with ample time available for systematic evaluation of the final competing candidates. In the formation of managers' day-to-day impressions of employee performance, however, this may not be the case. Managers often function under conditions of extreme time constraint and high cognitive load. In these situations, correction of stereotype-influenced impressions is far less likely to occur (Bargh, 1999; Gilbert, Pelham, and Krull, 1988).

To correct biased impressions, decision makers need ample decision-relevant information and access to structured decision-making processes. Carefully spelling out the applicable evaluative criteria, providing decision makers with objective, criterion-relevant information, and requiring them to write down how the information provided relates to the relevant criteria can help reduce the biasing effect of stereotypes on evaluations (Nieva and Gutek, 1980). As described earlier, providing decision makers with system-wide information revealing broad patterns of outcomes can help them identify biases they probably would not recognize from case-by-case data (Crosby et al., 1986).

Even though the implicit stereotypic expectancies do spontaneously bias initial impression formation, that bias is, at least under certain conditions, amenable to control. Organizations interested in maximizing the likelihood that correction will occur should implement the following strategies:

- Structure incentives and foster an organizational culture that motivates control whereby decision makers identify and correct implicitly biased impressions or initial judgments
- Monitor decisions systematically, so that broad patterns suggestive of uncorrected bias can be identified and addressed

People's need for control goes beyond mere rewards and punishments. People feel an intrinsic satisfaction in control for its own sake, which has been called an *effectance motive* (White, 1959). If increasing diversity is defined as an organizational goal, then discovering its contingencies and being effective in this way will have intrinsic reward without necessarily pinning specific performance criteria on this basis.

Harnessing the Motives for Self-Enhancement

Americans in particular, but people in other cultures in their own ways as well, are motivated to value the self. Americans are biased toward seeing themselves in a positive light, and they work to view themselves as both better than other people and better than other people see them (Kwan et al., 2004; Taylor and Brown, 1988).

Threats to one's self-esteem accordingly cause defensive action. In intergroup settings, retaliation often scapegoats the vulnerable groups (Fein and Spencer, 1997). When people discriminate against others, they temporarily feel better (Rubin and Hewstone, 1998). Accordingly, the opposite should hold: a diminishing threat and increasing stability and security all together allow people to interact better with others who differ from them.

Consider how this could apply:

- Destigmatize implicit bias so that self-examination and conscious correction are less threatening

Harnessing the Motives for Trust

People on the whole are motivated for social trust: they expect positive outcomes from others, at least from in-group others. This optimism means that people can be expected to have a baseline positivity bias, although they will be alert to negative information. Such baseline trust operates only with in-group others, those with whom one expects continuing ties. To make the most of this trust motive, the in-group must expand to include members of underrepresented groups, and all the recommendations for harnessing the belonging motive apply here.

In addition, consider how organizations might harness the motive to trust others:

- Emphasize an optimistic, positive perspective; promoting positive outcomes for the organization as a larger in-group should work better than avoiding negative outcomes, such as personal liability
- Emphasize cross-group enthusiasm, not mere tolerance

ORGANIZATIONAL INITIATIVES: SUMMARY AND CAVEAT

We have described a number of initiatives that organizations might undertake in an attempt to reduce the levels of workplace discrimination caused by unexamined, subtle bias. We end this discussion, however, with a caution. Very little research illuminates the critical question of what works, and what does not, in initiatives of this kind. We propose a research agenda for harnessing the social psychological principles of motivation and cognition.

To date, the most significant work on the subject is a study by Kalev, Dobbin, and Kelly (2006). Merging the federal EEO-1 database describing the workforces of 708 private companies from 1971 to 2002 with survey data from those same companies describing their EEO practices, Kalev and associates demonstrated that many of the best practices emerging from such agencies as the EEOC, the Federal Glass Ceiling Commission, and the Society for Human Resource Management actually had no significant impact on patterns of racial or gender diversity in the studied organizations. What mattered was accountability. Along these lines, the strongest predictors of levels of female and minority representation in management were (1) as noted, whether the company had been subjected to a compliance review by the OFCCP during a period of vigorous regulatory enforcement in the 1970s, and also (2) whether the company had established internal accountability systems for minority and female advancement (Kalev and Dobbin, 2006; Kalev, Kelly, and Dobbin, 2006). This work illustrates the necessity of a research agenda at the level of organizational policy.

Regulatory Enforcement Policy and Information Disclosure: Harnessing the Power of Accountability

Kalev, Kelly, and Dobbin's work is so important because it shows that many of the interventions that policy makers might *assume* reduce the expression of bias in the workplace (i.e., diversity training, policies against discrimination and harassment, etc.) actually have little, if any, effect. Apparently, what matters most is accountability for results. For employers to be held accountable for results, employees (current and prospective), regulators, and rights entrepreneurs need information about employers' EEO performance.

This information is, at present, almost completely unavailable. In this section, after describing the nature and extent of this problem, we will argue that regulatory initiatives requiring public disclosure of employers' EEO compliance records would increase the effectiveness of civil rights enforcement in a number of important ways.

ORGANIZATIONAL EQUAL EMPLOYMENT OPPORTUNITY COMPLIANCE AND THE VEIL OF SECRECY

In 2005, the EEOC received more than 23,000 charges of sex discrimination in employment. Of these, approximately 5,700 were resolved through voluntary settlement involving some remedy to the charging party. Just under 1,700 led to a formal administrative finding of discrimination by the EEOC (U.S. EEOC, 2006).

Under Section 706(b) of Title VII, however, the EEOC is prohibited from making public any information that would identify the employers, labor unions, or employment agencies against whom those charges of discrimination were filed or formally found. Any EEOC employee who makes such information public is guilty of a federal crime carrying a maximum prison sentence of up to one year.

The same confidentiality mandate cloaks other important EEO compliance information. For example, under Section 709(c) of Title VII, all private Title-VII employers with 100 or more employees, and all federal contractors who have 50 or more employees, must file annual EEO-1 reports that specify the proportion of women (and minorities) employed in each of nine job categories. Multifacility employers must file separate reports for each facility that employs 50 or more employees. These reports yield aggregated data from which patterns of underutilization by particular employers could readily be identified. However, under Section 709(e), no information from EEO-1 reports may be publicly disclosed in any way that would identify patterns of gender-, race-, or national-origin-related underutilization by individual companies.

Similar disclosure restrictions govern the EEO compliance program administered by the Department of Labor's OFCCP. Every federal contractor that employs 50 or more employees and has contracts worth over $50,000 with the federal government must, each year, file EEO compliance reports with the OFCCP. These reports contain a great deal of information, far more extensive than that provided by EEO-1s, from which prospective employees, current employees,

EEO advocacy groups, state EEO compliance agencies, and the EEOC could assess a company's EEO performance.

However, the OFCCP currently refuses to disclose to the public the affirmative action plans, compliance-review-related submissions, or annual Equal Opportunity Surveys it obtains from contractor employers. Currently, it takes the position that these materials are exempt from disclosure under the federal Freedom of Information Act (OFCCP, 2000). Presently, the OFCCP and the EEOC do not even share information with each other (Krieger, 2007).

Employers are also finding ways to keep out of the public sphere the kinds of EEO-compliance-related information that would ordinarily surface in discrimination lawsuits. As numerous legal commentators have observed, secrecy in American civil litigation is no longer the exception but the rule (Brenowitz, 2004; Doré, 2004). This secrecy is accomplished through a number of specific devices, including mandatory, predispute arbitration agreements imposed by employers as a condition of employment, confidentiality provisions in settlement agreements, and civil-discovery-related protective orders.

Many employers now require applicants, as a condition of employment, to agree that they will resolve any employment disputes, including claims of sex discrimination, through private arbitration rather than through access to the court system. In addition to depriving employees of a right to a jury trial, these mandatory arbitration arrangements allow employers to keep the existence and resolution of sex discrimination claims out of the public eye. Toward this end, many mandatory predispute arbitration agreements require employees to promise that they will maintain confidentiality, both during the arbitration and after its conclusion.

Second, almost universally now, when individual discrimination lawsuits settle, either before a case is actually filed or after it is filed but before a verdict is rendered, that settlement is conditioned on the employee/plaintiff's promise not to disclose the terms of the settlement. Given that over 95% of civil cases settle before trial (Gross and Syverud, 1996), the impact of routine confidentiality agreements in settlements can hardly be overstated. Other than the occasional high-profile, high-value jury award or government consent decree, the public has virtually no access to information about the nature or amounts of settlements in sex discrimination lawsuits or to the identity of the organizations from which such recoveries have been made.

Finally, protective orders issued by judges during the course of civil lawsuits, including high-profile class actions, are also increasingly used by employers to keep important evidence reflecting noncompliance with EEO laws out of the public view. The centerpiece of any pattern-and-practice sex discrimination lawsuit, like the centerpiece of an OFCCP compliance report, is a utilization analysis that compares the representation of women in the at-issue jobs with their availability in the qualified, relevant labor market. To prove discrimination, class-action plaintiffs often supplement this utilization analysis with evidence of other discrimination charges or findings against the employer, or with employer records or statements reflecting gender bias on the part of the decision makers.

The information from which a sex discrimination plaintiff can prove her case is generally obtained from the employer during a part of the lawsuit called discovery. Through the discovery process, employers are legally required to provide statistical and other information from which the plaintiff's lawyers—and ultimately the court—can assess whether the employer is systematically treating women less favorably than men and, if so, whether that difference in treatment can reasonably be attributed to gender bias. Absent a court order specifically prohibiting disclosure, once an employer provides this information to a sex discrimination plaintiff, the information can be shared with the public, with advocacy organizations, or with other aggrieved individuals and their counsel.

Increasingly, however, defendants in sex discrimination and other civil cases are obtaining broad protective orders through which the court prohibits plaintiffs from disclosing EEO-compliance-related information obtained during the discovery process. In their strongest form, these protective orders can require not only promises not to disclose information obtained in discovery (violations of which can be punished as contempt of court), but also the sealing of testimony, pleadings, exhibits, court transcripts, and other lawsuit-related documents, and the return of those materials to the employer after the litigation concluded, often by confidential settlement agreement.

These three devices—mandatory arbitration agreements with confidentiality provisions imposed as a condition of hiring, confidentiality clauses in settlement agreements, and the issuance of protective orders prohibiting disclosure of EEO-compliance-related information obtained in discovery—combine with EEOC and OFCCP confidentiality provisions to keep from the public virtually all employer-specific, systematically aggregated EEO compliance data. Neither individual women in the labor market nor the advocacy organizations that advance their interests have any way of knowing which employers are systematically treating women less favorably than men and which are not.

INFORMATION ASYMMETRIES AS A SOURCE OF INEFFICIENCY IN EQUAL EMPLOYMENT OPPORTUNITY COMPLIANCE POLICY

The almost total dearth of public information about employers' and union's EEO compliance records creates serious regulatory failures, which could be mitigated by regulation mandating greater transparency. Individual companies know (or at least have the capacity to know) whether they are utilizing women in proportion to their representation in the qualified, relevant selection pool. Employers also know, or can determine whether, after regressing out other relevant, nondiscriminatory independent variables, gender remains a significant predictor of compensation levels. Employers also know—or can easily determine—whether they have been subject to high levels of EEO complaints, charges, or lawsuits or whether particular practices are systematically functioning as obstacles to women's advancement within the organization.

But neither women contemplating new or continued employment with a particular firm nor women who think they may have been discriminated against in hiring, compensation, or promotion nor legal advocacy organizations nor even government compliance agencies have access to information of this sort. This creates two interrelated problems, one confronting prospective employees—"buyers" of employment opportunities, as it were—and one confronting regulators and "rights entrepreneurs," the legal advocacy organizations that seek to prevent sex discrimination from occurring or to redress it after the fact through litigation or other enforcement efforts.

Consider the first of these problems, the buyer's dilemma. As described earlier in this chapter, people have a great deal of difficulty identifying discrimination from individual, as opposed to aggregated, data. The unavailability to the public of employer- and union-specific EEO compliance data makes it virtually impossible for women in the labor market to know which employers are more or less likely to discriminate against them. Even if we assume, simply for purposes of argument, that market forces can eliminate some discrimination from the labor market, such forces cannot function in the face of such dramatic information asymmetries between the "sellers" of equal employment opportunities (employers) and the buyers of equal employment opportunity (prospective employees or current employees hoping to advance in an organization).

The information asymmetries also cause enforcement problems on the "back end" of a discrimination sequence. Because rights entrepreneurs lack access to meaningful EEO compliance information, neither legal advocacy organizations nor individual grievants can assess, before a lawsuit is filed and significant discovery is completed, the strength of a discrimination case. Thus, even if we do not believe that publicly available EEO compliance information would be optimally utilized by individual workers, its absence makes enforcement efforts substantially less effective and efficient.

And finally, given that employers can claim to support equal opportunity while hiding their actual compliance records behind a veil of secrecy, it is unlikely that "norm cascades" of the type described by Cass Sunstein (2006) could exert significant downward pressure on the rates of discrimination. Employers who engage in sex discrimination will not incur as many costs when the information that could identify them is publicly unavailable. For this additional reason, we suggest, mandatory information disclosure can be combined with legal prohibitions to encourage employers to overcome the tidal pulls of subtle bias and to act affirmatively to advance women into management and other high-level positions.

In sum, we suggest that to reckon successfully with subtle forms of gender bias, our EEO policy should decenter its current reliance on individual, after-the-fact disputes about whether discrimination occurred in any one particular case and rely more heavily on policy approaches that more systematically harness the power of accountability. This would entail supplementing the presently available set of EEO policy enforcement tools with a system of mandatory, public information disclosure requirements, similar to those now used to regulate the securities markets. Admittedly, this policy prescription is based largely on basic, and still contested, ideas drawn from information economics, in which the identification and elimination of information asymmetries is mobilized to make markets more efficient (Stiglitz, 2000).

Our use of insights from information economics, however, comes with a twist. Conventional economic analysis of the roles of information asymmetries, signaling, and screening has uniformly cast prospective employees as sellers, who must signal their performance potential to employers, who function in those analyses as buyers and screeners (Cooter, 1994; Spence, 1973; Stiglitz, 2002). We suggest here, however, that to reckon effectively with subtle forms of gender bias in the labor market, this conception of the roles of signaling and screening must be turned on its head. Instead of asking only what information employees have that employers need, economic analysis of employment discrimination must also ask what information employers have about their own EEO performance that current or potential employees need to make rational, utility-maximizing decisions about their own career choices, such as

- Utilization analyses comparing the percentage of women in each job group with their availability in the relevant selection pool
- An explanation of how, for each job group, the relevant selection-pool statistics were compiled, with the explanation presented in sufficient detail that a reasonably sophisticated reader could evaluate whether the availability pool and job groups had been properly constructed
- Statistics showing, for each job group and at each level up the relevant line of progression, the comparative selection rates and odds ratios of women versus men being promoted or otherwise advanced from one level or step to the next
- For each compensation classification, statistics showing the average wage or salary gap, if any, between male and female employees in a particular job group at each step up the job ladder
- For each job group in which women are underutilized relative to their representation in the available selection pool, an analysis of the current obstacles to women's inclusion or advancement and, if applicable, the employer, union, or employment agency's plan for removing them
- Statistics disclosing on a facility-by-facility basis the number of sex discrimination charges filed with either the EEOC, the OFCCP, or any other EEO enforcement agency (state or federal), with a description of the issue alleged in each complaint (i.e. promotion, compensation, harassment, termination, etc.) and an indication of whether, and if so, how, the complaint was resolved

Summary of Modern Policy Tools

We have outlined steps that organizations can themselves undertake to reduce the expression of subtle gender bias in employment decision making, and we proposed sweeping changes to the federal government's approach to enforcing EEO laws and regulations. In doing so, we have revisited the question of how market forces can be mobilized to squeeze sex discrimination out of substantial segments of the labor market, and in this regard, have recommended the institution of a comprehensive, mandatory information disclosure regime similar to that characterizing the regulation of U.S. securities markets.

Conclusion

As we have shown, subtle, modern forms of bias are automatic, ambiguous, and ambivalent. This makes their impact on particular employment decisions particularly difficult and costly to identify. If implicit bias increasingly plays a relatively larger role than overt forms in limiting equal employment opportunity for women, minorities, and other protected groups, society would be drawn increasingly to systemic, structural approaches to EEO compliance policy and away from individual adjudications as our primary policy tool.

Note

1. We do not suggest that the private right of action to sue for individual disparate treatment discrimination be eliminated. However, the second author has in another context argued that in cases involving discrimination caused by the unwitting application of implicit stereotypes, only equitable remedies, and not compensatory or punitive damages, be available to prevailing plaintiffs (Krieger, 1995).

References

Akerlof, G. A. (1970). The market for "lemons:" Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, *84*, 488–500.

Alexander, M. G., Brewer, M. B., and Hermann, R. K. (1999). Images and affect: A functional analysis of out-group stereotypes. *Journal of Personality and Social Psychology*, *77*, 78–93.

American Association of University Women (2007). *Behind the pay gap*. Washington, DC: AAUW Educational Foundation.

Amodio, D. M. (2008). The social neuroscience of intergroup relations. *European Review of Social Psychology*, *19*, 1–54.

Ashburn-Nardo, L., Voils, C. I., and Monteith, M. J. (2001). Implicit associations as the seeds of intergroup bias: How easily do they take root? *Journal of Personality and Social Psychology*, *81*, 789–799.

Babcock, L., and Laschever, S. (2003). *Women don't ask: Negotiation and the gender divide*. Princeton, NJ: Princeton University Press.

Bagenstos, S. R. (2006). The structural turn and the limits of antidiscrimination law. *California Law Review*, *94*, 1–47.

Banaji, M. R., and Hardin, C. D. (1996). Automatic stereotyping. *Psychological Science*, *7*, 136–141.

Banaji, M. R., Hardin, C., and Rothman, A. J. (1993). Implicit stereotyping in person judgment. *Journal of Personality and Social Psychology*, *65*, 272–281.

Bargh, J. A. (1999). The cognitive monster: The case against the controllability of automatic stereotype effects. In S. Chaiken and Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 361–382). New York: Guilford

Bargh, J. A., Raymond, P., Pryor, J. B., and Strack, F. (1995). Attractiveness of the underling: An automatic power-sex association and its consequences for sexual harassment and aggression. *Journal of Personality and Social Psychology*, *68*, 768–781.

Belkin, L. (2003, October 26). The opt-out revolution. *New York Times Magazine*, pp. 42–46.

———. (2005, October 9). In-house nepotism: Hiring the husband. *New York Times Job Market*, Section 10, p. 1.

Berdahl, J. L. (2007). Harassment based on sex: Protecting social status in the context of gender hierarchy. *Academy of Management Review*, *32*, 641–658.

Biernat, M., Crandall, C. S., Young, L. V., Kobrynowicz, D., and Halpin, S. M. (1998). All that you can be: Stereotyping of self and others in a military context. *Journal of Personality and Social Psychology*, *75*, 301–317.

Biernat, M., and Fuegen, K. (2001). Shifting standards and the evaluation of competence: Complexity in gender-based judgment and decision making. *Journal of Social Issues*, *57*, 707–724.

Biernat, M., and Kobrynowicz, D. (1997). Gender- and race-based standards of competence: Lower minimum standards but higher ability standards for devalued groups. *Journal of Personality and Social Psychology*, *72*, 544–557.

Biernat, M., and Vescio, T. K. (2002). She swings, she hits, she's great, she's benched: Implications of gender-based shifting standards for judgment and behavior. *Personality and Social Psychology Bulletin*, *28*, 66–77.

Blair, I. V., and Banaji, M. R. (1996). Automatic and controlled processes in stereotype priming. *Journal of Personality and Social Psychology*, *70*, 1142–1163.

Blair, I. V., Ma, J. E., and Lenton, A. P. (2001). Imagining stereotypes away: The moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology*, *81*, 828–841.

Blau, F. D., Brinton, M. C., and Grusky, D. (Eds.) (2006). *The declining significance of gender?* New York: Russell Sage Foundation.

Bodenhausen, G. A., and Lichtenstein, M. (1987). Social stereotypes and information-processing strategies: The impact of task complexity. *Journal of Personality and Social Psychology*, *52*, 871–880.

Bodenhausen, G. A., and Wyer, R. S. (1985). Effects of stereotypes in decision making and information-processing strategies. *Journal of Personality and Social Psychology*, *48*, 267–282.

Brenowitz, S. (2004). Deadly secrecy: The erosion of public information under private justice. *Ohio State Journal on Dispute Resolution, 19*, 679–708.

Brewer, M. B. (1996). In-group favoritism: the subtle side of intergroup discrimination. In D. M. Messick and A. E. Tenbrunsel (Eds.), *Codes of conduct: Behavioral research and business ethics* (pp. 160–71). New York: Russell Sage Foundation.

———. (2000). Reducing prejudice through cross-categorization: Effects of multiple social identities. In S. Oskamp (Ed.), *Reducing prejudice and discrimination* (pp. 165–184). Mahwah, NJ: Erlbaum.

Carvallo, M., and Pelham, B. W. (2006). When fiends becomes friends: The need to belong and perceptions of personal and group discrimination. *Journal of Personality and Social Psychology*, *90*, 94–108.

Castelli, L., Macrae, C. N., Zogmaister, C., and Arcuri, L. (2004). A tale of two primes: Contextual limits on stereotype activation. *Social Cognition*, *22*, 233–247.

Catalyst (2002). *2002 Catalyst census of women corporate officers and top earners in the Fortune 500*. Retrieved from http://www.catalyst.org/file/87/2002%20catalyst%20census%20of%20women%20corporate%20officers%20and%20top%20earners%20of%20the%20fortune%20500.pdf

Chaiken, S., and Trope, Y. (Eds.). (1999). *Dual-process theories in social psychology*. New York: Guilford.

Charny, D. and Gulati, G. M. (1998). Efficiency wages, tournaments, and discrimination: A theory of employment discrimination law for "high level" jobs. *Harvard Civil Rights–Civil Liberties Law Review*, *33*, 57–105.

Clausell, E., and Fiske, S. T. (2005). When do the parts add up to the whole? Ambivalent stereotype content for gay male subgroups. *Social Cognition*, *23*, 157–176.

Cloutier, J., Mason, M. F., and Macrae, C. N. (2005). The perceptual determinants of person construal: Reopening the social-cognitive toolbox. *Journal of Personality and Social Psychology*, *88*, 885–894.

Cooter, R. (1994). Market affirmative action. *San Diego Law Review*, *31*, 133–168.

Crocker, J., and Major, B. (1989). Social stigma and self-esteem: The self-protective properties of stigma. *Psychological Review*, *96*, 608–630.

Crocker, J., Voelkl, K., Testa, M., and Major, B. (1991). Social stigma: The affective consequences of attributional ambiguity. *Journal of Personality and Social Psychology*, *60*, 218–228.

Crosby, F. (1984). The denial of personal discrimination. *American Behavioral Scientist*, *27*, 371–386.

Crosby, F., Clayton, S., Alksnis, O., and Hemker, K. (1986). Cognitive biases in the perception of discrimination: The importance of format. *Sex Roles*, *14*, 637–646.

Cuddy, A.J.C., Fiske, S. T., and Glick, P. (2007). The BIAS map: Behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology*, *92*, 631–648.

Cuddy, A.J.C., Fiske, S. T., Kwan, V.S.Y., Glick, P., Demoulin, S., Leyens, J-Ph., et al. (2009). Is the stereotype content model culture-bound? A cross-cultural comparison reveals systematic similarities and differences. *British Journal of Social Psychology*, *48*, 1–33.

Cunningham, W. A., Preacher, K. J., and Banaji, M. R. (2001). Implicit attitude measures: Consistency,

stability, and convergent validity. *Psychological Science*, *12*, 163–170.

Czopp, A. M., and Monteith, M. J. (2003). Confronting prejudice (literally): Reactions to confrontations of racial and gender bias. *Personality and Social Psychology Bulletin*, *29*, 532–544.

Dasgupta, N. (2004). Implicit ingroup favoritism, outgroup favoritism, and their behavioral manifestations. *Social Justice Research*, *17*, 143–169.

Deaux, K., and Emswiller, E. (1974). Explanations of successful performance on sex-linked tasks: What is skill for the male is luck for the female. *Journal of Personality and Social Psychology*, *29*, 80–85.

Diekman, A. B., and Eagly, A. H. (2000). Stereotypes as dynamic constructs: Women and men of the past, present, and future. *Personality and Social Psychology Bulletin*, *26*, 1171–1188.

Dijksterhuis, A., and van Knippenberg, A. (2000). Behavioral indecision: Effects of self-focus on automatic behavior. *Social Cognition*, *18*, 55–74.

Donohue, J. J., III. (1986). Is Title VII efficient? *University of Pennsylvania Law Review*, *134*, 1411–1431.

———. (1987). Further thoughts on employment discrimination legislation: A reply to Judge Posner. *University of Pennsylvania Law Review*, *136*, 523–551.

———. (1989). Prohibiting sex discrimination in the workplace: An economic perspective. *University of Chicago Law Review*, *56*, 1337–1368.

Donohue, J. J., III, and Siegelman, P. (1991). The changing nature of employment discrimination litigation. *Stanford Law Review*, *43*, 983–1030.

Doré, L. K. (2004). Settlement, secrecy, and judicial discretion: South Carolina's new rules governing the sealing of settlements. *South Carolina Law Review*, *55*, 791–827.

Dovidio, J. F., Evans, N., and Tyler, R. B. (1986). Racial stereotypes: The contents of their cognitive representations. *Journal of Experimental Social Psychology*, *22*, 22–37.

Dovidio, J. F., Kawakami, K., and Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, *82*, 62–68.

Dunning, D., and Sherman, D. A. (1997). Stereotypes and tacit inference. *Journal of Personality and Social Psychology*, *73*, 459–471.

Eagly, A. H., and Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, *109*, 573–598.

Eagly, A. H., and Steffen, V. J. (1984). Gender stereotypes stem from the distribution of women and men into social roles. *Journal of Personality and Social Psychology*, *46*, 735–754.

Eckes, T. (2002). Paternalistic and envious gender stereotypes: Testing predictions from the stereotype content model. *Sex Roles*, *47*, 99–114.

Epstein, R. A. (1995). *Forbidden grounds: The case against employment discrimination laws*. Cambridge, MA: Harvard University Press.

Estlund, C. L. (2000). Working together: The workplace, civil society, and the law. *Georgetown Law Journal*, *89*, 1–96.

Faludi, S. (1991). *The undeclared war against American women*. New York: Crown.

Fazio, R. H., and Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, *54*, 297–327.

Fein, S., and Spencer, S. J. (1997). Prejudice as self-image maintenance: Affirming the self through derogating others. *Journal of Personality and Social Psychology*, *73*, 31–44.

Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. In D. T. Gilbert, S. T. Fiske, and G. Lindzey (Eds.), *Handbook of social psychology* (4th ed., Vol. 2, pp. 357–411). New York: McGraw-Hill.

———. (2007). Core social motivations, a historical perspective: Views from the couch, consciousness, classroom, computers, and collectives. In W. Gardner and J. Shah (Eds.), *Handbook of motivation science* (pp. 3–22). New York: Guilford.

———. (2010a). *Social beings: A core motives approach to social psychology*. New York: Wiley.

———. (2010b). Stratification. In S. T. Fiske, D. T., Gilbert, and G. Lindzey (Eds.) *Handbook of social psychology* (5th ed., pp. 941–982). New York: Wiley.

Fiske, S. T., Bergsieker, H., Russell, A. M., and Williams, L. (2009). Images of Black Americans: Then, "them" and now, "Obama!" *DuBois Review: Social Science Research on Race*, *6*, 83–101.

Fiske, S. T., Bersoff, D. N., Borgida, E., Deaux, K., and Heilman, M. E. (1991). Social science research on trial: The use of sex stereotyping research in *Price Waterhouse v. Hopkins*. *American Psychologist*, *46*, 1049–1060.

Fiske, S. T., Cuddy, A.J.C., and Glick, P. (2007). Universal dimensions of social perception: Warmth and competence. *Trends in Cognitive Science*, *11*, 77–83.

Fiske, S. T., Lin, M. H., and Neuberg, S. L. (1999). The continuum model: Ten years later. In S. Chaiken and Y. Trope (Eds.) *Dual process theories in social psychology* (pp. 231–254). New York: Guilford.

Fiske, S. T., and Taylor, S. E. (2008). *Social cognition: From brains to culture*. New York: McGraw-Hill.

Frantz, C. M., Cuddy, A.J.C., Burnett, M., Ray, H., and Hart, A. (2004). A threat in the computer: The race implicit association test as a stereotype threat experience. *Personality and Social Psychology Bulletin*, *30*, 1611–1624.

Gaertner, S. L., and Dovidio, J. F. (2005). Understanding and addressing contemporary racism: From aversive racism to the common ingroup identity model. *Journal of Social Issues*, *61*, 615–639.

Gaertner, S. L., and McLaughlin, J. P. (1983). Racial stereotypes: Associations and ascriptions of positive and negative characteristics. *Social Psychology Quarterly*, *46*, 23–30.

Gaertner, S. L., Sedikides, C., and Graetz, K. (1999). In search of self-definition: Motivational primacy of the individual self, motivational primacy of the collective self, or contextual primacy? *Journal of Personality and Social Psychology*, *76*, 5–18.

Galinsky, A. D., and Moskowitz, G. B. (2000). Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of Personality and Social Psychology*, *78*, 708–724.

Gallup Organization (2005). *Employee discrimination in the workplace*. Washington, DC: Gallup. Retrieved from http://media.gallup.com/government/PDF/Gallup_Discrimination_Report_Final.pdf

Gilbert, D. T., Pelham, B. W., and Krull, D. S. (1988). On cognitive busyness: When person perceivers meet persons perceived. *Journal of Personality and Social Psychology*, *54*, 733–740.

Glick, P., Diebold, J., Bailey-Werner, B., and Zhu, L. (1997). The two faces of Adam: Ambivalent sexism and polarized attitudes toward women. *Personality and Social Psychology Bulletin*, *23*, 1323–1334.

Glick, P., and Fiske, S. T. (1996). The Ambivalent Sexism Inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, *70*, 491–512.

———. (2001). Ambivalent sexism. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 33, pp. 115–188). San Diego, CA: Academic Press.

Glick, P., Fiske, S. T., Mladinic, A., Saiz, J. L., Abrams, D., Masser, B., et al. (2000). Beyond prejudice as simple antipathy: Hostile and benevolent sexism across cultures. *Journal of Personality and Social Psychology*, *79*, 763–775.

Glick, P., Thomas, M., Vescio, T., and Fiske, S. T. (In preparation). *The stereotype-confirming attributional bias*. Manuscript in preparation.

Goldberg, P. (1968). Are women prejudiced against women? *Transaction*, *5*, 28–30.

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., and Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, *109*, 3–25.

Greenwald, A. G., McGhee, D. E., and Schwartz, J.L.K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology Bulletin*, *74*, 1464–1480.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., and Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*, 17–41.

Gruenfeld, D. H., and Tiedens, L. Z. (2010). On the social psychology of organizing. In S. T. Fiske, D. T. Gilbert, and G. Lindzey (Eds.), *Handbook of social psychology* (5th ed., pp. 1252–1287). New York: Wiley.

Gross, S. R., and Syverud, K. D. (1996). Don't try: Civil jury verdicts in a system geared to settlement. *UCLA Law Review*, *4*, 1–64.

Heilman, M. E. (1983). Sex bias in work settings: The Lack of Fit Model. *Research in Organizational Behavior*, *5*, 269–298.

Heilman, M. E., and Haynes, M. C. (2008). Subjectivity in the appraisal process: A facilitator of gender bias in work settings. In E. Borgida and S. T. Fiske (Eds.), *Psychological science in the courtroom: Beyond common sense* (pp. 127–156). London: Blackwell.

Hewstone, M. (1990). The "ultimate attribution error": A review of the literature on intergroup causal attribution. *European Journal of Social Psychology*, *20*, 311–335.

Hoffman, C., and Hurst, N. (1990). Gender stereotypes: Perception or rationalization? *Journal of Personality and Social Psychology*, *58*, 197–208.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., and Schmitt, M. (2005). A meta-analysis on the correlation between the implicit association test and explicit self-report measures. *Personality and Social Psychology Bulletin*, *31*, 1369–1385.

Hopkins v. Price Waterhouse, 618 F. Supp. 1109 (D.D.C. 1985); appeal: 825 F.2d 458 (D.C. Cir. 1987); Supreme Court review: 109 S. Ct. 1775 (1989); remand: No. 84-3040, slip op. (D.D.C. May 14, 1990).

Ito, T. A., and Urland, G. R. (2003). Race and gender on the brain: Electrocortical measures of attention to the race and gender of multiply categorizable individuals. *Journal of Personality and Social Psychology*, *85*, 616–626.

Kaiser, C. R., and Miller, C. T. (2001). Stop complaining! The social costs of making attributions to discrimination. *Personality and Social Psychology Bulletin*, *27*, 254–263.

———. (2003). Derogating the victim: The interpersonal consequences of blaming events on discrimination. *Group Processes and Intergroup Relations*, *6*, 227–237.

Kalev, A. and Dobbin, F. (2006). Enforcement of civil rights law in private workplaces: The effects of compliance reviews and lawsuits over time. *Law and Social Inquiry*, *31*, 855–903.

Kalev, A., Kelly, E. and Dobbin, F. (2006). Best practices or best guesses? Assessing the efficacy of corporate affirmative action and diversity policies. *American Sociological Review*, *71*, 589–617.

Kang, J. (2005). Trojan horses of race. *Harvard Law Review*, *118*, 1489–1593.

Katz, I., and Hass, R. G. (1988). Racial ambivalence and American value conflict: Correlational and priming studies of dual cognitive structures. *Journal of Personality and Social Psychology*, *55*, 893–905.

Katz, I., Wackenhut, J., and Hass, R. G. (1986). Racial ambivalence, value duality, and behavior. In J. F. Dovidio and S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 35–59). San Diego, CA: Academic Press.

Kawakami, K., and Dovidio, J. F. (2001). The reliability of implicit stereotyping. *Personality and Social Psychology Bulletin*, *27*, 212–225.

Kawakami, K., Dovidio, J. F., Moll, J., Hermsen, S., and Russin, A. (2000). Just say no (to stereotyping): Effects of training in the negation of stereotypic associations on stereotype activation. *Journal of Personality and Social Psychology*, *78*, 871–888.

Kawakami, K., Dovidio, J. F., and van Kamp, S. (2005). Kicking the habit: Effects of nonstereotypic association training and correction processes on hiring decisions. *Journal of Experimental Social Psychology*, *41*, 68–75.

Klauer, K. C., and Wegener, I. (1998). Unraveling social categorization in the "Who said what?" paradigm. *Journal of Personality and Social Psychology*, *75*, 1155–1178.

Krieger, L. H. (1995). The content of our categories: A cognitive bias approach to discrimination and equal employment opportunity. *Stanford Law Review*, *47*, 1161–1247.

———. (1998). Civil rights perestroika: Intergroup relations after affirmative action. *California Law Review*, *86*, 1251–1333.

———. (2007). The watched variable improves: On eliminating sex discrimination in employment. In F. Crosby, M. Stockdale, and S. A. Ropp (Eds.), *Sex discrimination in the workplace (*pp. 295–331). Malden: Blackwell.

Krieger, L. H. and Fiske, S. T. (2006). Behavioral realism in employment discrimination law: Implicit bias and disparate treatment. *University of California Law Review*, *94*, 997–1062.

Kunda, Z., and Sherman-Williams, B. (1993). Stereotypes and the construal of individuating information. *Personality and Social Psychology Bulletin*, *19*, 90–99.

Kwan, V.S.Y., John, O. P., Kenny, D. A., Bond, M. H., and Robins, R. W. (2004). Reconceptualizing individual differences in self-enhancement bias: An interpersonal approach. *Psychological Review*, *111*, 94–110.

Lemm, K. M., Dabady, M., and Banaji, M. R. (2005). Gender picture priming: It works with denotative and connotative primes. *Social Cognition*, *23*, 218–241.

Leonhardt, D. (December 24, 2006). Gender pay gap, once narrowing, is stuck in place. *New York Times.* Retrieved from http://www.nytimes.com/2006/12/24/business/24gap.html

Lieberman, M. D., Gaunt, R., Gilbert, D. T., and Trope, Y. (2002). Reflection and reflexion: A social cognitive neuroscience approach to attributional inference. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 34, pp. 199–249). San Diego, CA: Academic Press.

Lord, C. G., Lepper, M. R., and Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology*, *47*, 1231–1243.

Lyness, K. S., and Judiesch, M. K. (1999) Are women more likely to be hired or promoted into management positions? *Journal of Vocational Behavior*, *54*, 158–173.

Macrae, C. N., Bodenhausen, G. V., and Milne, A. B. (1998). Saying no to unwanted thoughts: Self-focus and the regulation of mental life. *Journal of Personality and Social Psychology*, *74*, 578–589.

Macrae, C. N., Bodenhausen, G. V., Milne, A. B., and Ford, R. L. (1997). On regulation of recollection: The intentional forgetting of stereotypical memories. *Journal of Personality and Social Psychology*, *72*, 709–719.

Macrae, C. N., Bodenhausen, G. V., Milne, A. B., and Jetten, J. (1994). Out of mind but back in sight: Stereotypes on the rebound. *Journal of Personality and Social Psychology*, *67*, 808–817.

Macrae, C. N., Hewstone, M., and Griffiths, R. J. (1993). Processing load and memory for stereotype-based information. *European Journal of Social Psychology*, *23*, 77–87.

Macrae, C. N., Milne, A. B., and Bodenhausen, G. V. (1994). Stereotypes as energy saving devices: A peek inside the cognitive toolbox. *Journal of Personality and Social Psychology*, *66*, 37–47.

Macrae, C. N., Quinn, K. A., Mason, M. F., and Quadflieg, S. (2005). Understanding others: The face and person construal. *Journal of Personality and Social Psychology*, *89*, 686–695.

Maddox, K. B., and Chase, S. G. (2004). Manipulating subcategory salience: Exploring the link between skin tone and social perception of Blacks. *European Journal of Social Psychology*, *34*, 533–546.

Maddox, K. B., and Gray, S. A. (2002). Cognitive representations of Black Americans: Reexploring the role of skin tone. *Personality and Social Psychology Bulletin*, *28*, 250–259.

Major, B., Gramzow, R. H., McCoy, S. K., Levin, S., Schmader, T., and Sidanius, J. (2002). Perceiving personal discrimination: The role of group status and legitimizing ideology. *Journal of Personality and Social Psychology*, *82*, 269–282.

Major, B., Quinton, W. J., and McCoy, S. K. (2002). Antecedents and consequences of attributions to

discrimination: Theoretical and empirical advances. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 34, pp. 251–330). San Diego, CA: Academic Press.

Major, B., Testa, M., and Blysma, W. H. (1991). Responses to upward and downward social comparisons: The impact of esteem-relevance and perceived control. In J. Suls and T. A. Wills (Eds.), *Social comparison: Contemporary theory and research* (pp. 237–260). Hillsdale, NJ: Erlbaum.

Masser, B. M., and Abrams, D. (2004). Reinforcing the glass ceiling: The consequences of hostile sexism for female managerial candidates. *Sex Roles*, *51*, 609–615.

Massey, D. S. (2007). *Categorically unequal: The American stratification system*. New York: Russell Sage Foundation.

McAdams, R. H. (1995). Cooperation and conflict: The economics of group status production and race discrimination. *Harvard Law Review*, *108*, 1003–1084.

McConnell, A. R., and Fazio, R. H. (1996). Women as men and people: Effects of gender-marked language. *Personality and Social Psychology Bulletin*, *22*, 1004–1013.

McKenzie-Mohr, D., and Zanna, M. P. (1990). Treating women as sexual objects: Look to the (gender schematic) male who has viewed pornography. *Personality and Social Psychology Bulletin*, *16*, 296–308.

Mitchell, J. P., Nosek, B. A., and Banaji, M. R. (2003). Contextual variations in implicit evaluation. *Journal of Experimental Psychology: General*, *132*, 455–469.

Monin, B., and Miller, D. T. (2001). Moral credentials and the expression of prejudice. *Journal of Personality and Social Psychology*, *81*, 33–43.

Monteith, M. J., Sherman, J. W., and Devine, P. G. (1998). Suppression as a stereotype control strategy. *Personality and Social Psychology Review*, *2*, 63–82.

Monteith, M. J., Spicer, C. V., and Tooman, G. D. (1998). Consequences of stereotype suppression: Stereotypes on AND not on the rebound. *Journal of Experimental Social Psychology*, *34*, 355–377.

Monteith, M. J. and Voils, C. I. (2001). Exerting control over prejudiced responses. In G. B. Moskowitz (Ed.), *Cognitive social psychology: The Princeton symposium on the legacy and future of social cognition*. Mahwah, NJ: Erlbaum.

Murray, J. D., Spadafore, J. A., and McIntosh, W. D. (2005). Belief in a just world and social perception. *Journal of Social Psychology*, *145*, 35–47

Neilsen, L. B. and Nelson, R. L. (2005). Rights realized?: An empirical analysis of employment discrimination litigation as a claiming system. *Wisconsin Law Review*, 663–711.

Nieva, V. F. and Gutek, B. A. (1980). Sex effects on evaluation, *Academy of Management Review*, *5*, 267–276.

Norton, M. I., Vandello, J. A., and Darley, J. M. (2004). Casuistry and social category bias. *Journal of Personality and Social Psychology*, *87*, 817–831.

Nosek, B. A., Banaji, M. R., and Greenwald, A. G. (2002a). Harvesting implicit group attitudes and beliefs from a demonstration website. *Group Dynamics: Theory, Research, and Practice*, *6*, 101–115.

———. (2002b). Math = male, me = female, therefore math ≠ me. *Journal of Personality and Social Psychology*, *83*, 44–59.

O'Connor, M., Gutek, B. A., Stockdale, M., Geer, T. M., and Melançon, R. (2004). Explaining sexual harassment judgments: Looking beyond gender of the rater. *Law and Human Behavior*, *28*, 69–95.

Office of Federal Contract Compliance Programs (OFCCP), Equal Employment Opportunity, U. S. Dept. of Labor, Compliance Evaluations 41 C.F.R. 60-1.20, 60-2.18(d) (2000).

Perdue, C. W., Dovidio, J. F., Gurtman, M. B., and Tyler, R. B. (1990). Us and them: Social categorization and the process of intergroup bias. *Journal of Personality and Social Psychology*, *59*, 475–486.

Pettigrew, T. F. (1979). The ultimate attribution error: Extending Allport's cognitive analysis of prejudice. *Personality and Social Psychology Bulletin*, *5*, 461–476.

Phalet, K., and Poppe, E. (1997). Competence and morality dimensions of national and ethnic stereotypes: A study in six eastern-Europe countries. *European Journal of Social Psychology*, *27*, 703–723.

Pinel, E. C. (1999). Stigma consciousness: The psychological legacy of social stereotypes. *Journal of Personality and Social Psychology*, *76*, 114–128.

———. (2002). Stigma consciousness in intergroup contexts: The power of conviction. *Journal of Experimental Social Psychology*, *76*, 114–128.

Plant, E. A., Kling, K. C., and Smith, G. L. (2004). The influence of gender and social role on the interpretation of facial expressions. *Sex Roles*, *51*, 187–196.

Prentice, D. A., and Miller, D. T. (2006). Essentializing differences between women and men. *Psychological Science*, *17*, 129–135.

Posner, R. A. (1989). An economic analysis of sex discrimination laws. *University of Chicago Law Review*, *56*, 1311–1335.

———. (1987). The efficiency and efficacy of Title VII. *University of Pennsylvania Law Review*, *136*, 513–522.

Pronin, E., Lin, D. Y., and Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, *28*, 369–381.

Quinn, K. A., and Macrae, C. N. (2005). Categorizing others: The dynamics of person construal. *Journal of Personality and Social Psychology*, *88*, 467–479.

Reskin, I., and Padavic, B. (2002). *Women and men at work* (2nd ed.). Thousand Oaks, CA: Pine Forge Press.

Reskin, I., and Ross, C. E. (1992). Jobs, authority, and earnings among managers: The continuing significance of sex. *Work and Occupations*, *19*, 342–365.

Riach, P. A., and Rich, J. (2002). Field experiments of discrimination in the market place. *Economic Journal*, *112*, F480–F518.

Richeson, J. A., Baird, A. A., Gordon, H. L., Heatherton, T. F., Wyland, C. L., Trawalter, S., and Shelton, J. N. (2003). An fMRI investigation of the impact of interracial contact on executive function. *Nature Neuroscience*, *6*, 1323–1328.

Richeson, J. A., and Shelton, J. N. (2003). When prejudice does not pay: Effects of interracial contact on executive function. *Psychological Science*, *14*, 287–290.

Richeson, J. A., Trawalter, S., and Shelton, J. N. (2005). African American's implicit racial attitudes and the depletion of executive function after interracial interactions. *Social Cognition*, *23*, 336–352.

Roose, P. A., and Gatta, M. L. (1999). The gender gap in earnings: Trends, explanations, and prospects. In G. Powell, (Ed.), *Handbook of gender and work* (pp. 95–123). Thousand Oaks, CA: Sage.

Rubin, M., and Hewstone, M. (1998). Social identity theory's self-esteem hypothesis: A review and some suggestions for clarification. *Personality and Social Psychology Review*, *2*, 40–62.

Rudman, L. A., and Borgida, E. (1995). The afterglow of construct accessibility: The behavioral consequences of priming men to view women as sexual objects. *Journal of Experimental Social Psychology*, *31*, 493–517.

Rudman, L. A., and Glick, P. (1999). Feminized management and backlash toward agentic women: The hidden costs to women of a kinder, gentler image of middle-managers. *Journal of Personality and Social Psychology*, *77*, 1004–1010.

———. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues*, *57*, 743–762.

———. (2008). *Social psychology of gender: How power and intimacy shape gender relations*. New York: Guilford Press.

Rudman, L. A., and Goodwin, S.A. (2004). Gender differences in automatic in-group bias: Why do women like women more than men like men? *Journal of Personality and Social Psychology*, *87*, 494–509.

Rudman, L. A., Greenwald, A. G., and McGhee, D. E. (2001). Implicit self-concept and evaluative implicit gender stereotypes: Self and ingroup share desirable traits. *Personality and Social Psychology Bulletin*, *27*, 1164–1178.

Rudman, L. A., Greenwald, A .G., Mellott, D. S., and Schwartz, J.L.K. (1999). Measuring the automatic components of prejudice: Flexibility and generality of the Implicit Association Test. *Social Cognition*, *17*, 437–465.

Rudman, L. A., and Kilianski, S. E. (2000). Implicit and explicit attitudes toward female authority. *Personality and Social Psychology Bulletin*, *26*, 1315–1328.

Russell, B. L., and Trigg, K. Y. (2004). Tolerance of sexual harassment: An examination of gender differences, ambivalent sexism, social dominance, and gender roles. *Sex Roles*, *50*, 565–573.

Rutte, C. G., Diekmann, K. A., Polzer, J. T., Crosby, F. J., and Messick, D. M. (1994). Organization of information and the detection of gender discrimination. *Psychological Science*, *5*, 226–231.

Sailer, P., Yau, E., and Rehula, V. (2002). Income by gender and age from information returns. *Statistics of Income Bulletin*, *21*, 83–102.

Santuzzi, A. M., and Ruscher, J. B. (2002). Stigma salience and paranoid social cognition: Understanding variability in metaperceptions among individuals with recently-acquired stigma. *Social Cognition*, *20*, 171–198.

Schmitt, M. T., Branscombe, N. R., and Postmes, T. (2003). Women's emotional responses to the pervasiveness of gender discrimination. *European Journal of Social Psychology*, *33*, 297–312.

Shelton, J. N. (2003). Interpersonal concerns in social encounters between majority and minority group members. *Group Processes and Intergroup Relations*, *6*, 171–185.

Shelton, J. N., and Richeson, J. A. (2006). Interracial interactions: A relational approach. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 38, pp. 121–181). San Diego, CA: Academic Press.

Sherman, J. W., Conrey, F. R., and Groom, C. J. (2004). Encoding flexibility revisited: Evidence for enhanced encoding of stereotype-inconsistent information under cognitive load. *Social Cognition*, *22*, 214–232.

Sherman, J. W., Klein, S. B., Laskey, A., and Wyer, N. A. (1998). Intergroup bias in group judgment processes: The role of behavioral memories. *Journal of Experimental Social Psychology*, *34*, 51–65.

Sherman, J. W., Lee, A. Y., Bessenoff, G. R., and Frost, L. A. (1998). Stereotype efficiency reconsidered: Encoding flexibility under cognitive load. *Journal of Personality and Social Psychology*, *75*, 589–606.

Sherman, J. W., Stroessner, S. J., Conrey, F. R., and Azam, O. A. (2005). Prejudice and stereotype maintenance processes: Attention, attribution, and individuation. *Journal of Personality and Social Psychology*, *89*, 607–622.

Sinclair, S., Lowery, B. S., Hardin, C. D., and Colangelo, A. (2005). Social tuning of automatic racial attitudes: The role of affiliative motivation. *Journal of Personality and Social Psychology*, *89*, 583–592.

Son Hing, L. S., Li, W., and Zanna, M. P. (2002). Inducing hypocrisy to reduce prejudicial responses among aversive racists. *Journal of Experimental Social Psychology*, *38*, 71–78.

Spence, J. T., and Helmreich, R. L. (1972). The Attitudes toward Women Scale: An objective instrument to measure attitudes toward the rights and roles of women in contemporary society. *JSAS Catalog of Selected Documents in Psychology*, *2*, 66–67.

Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, *87*, 355–374.

Stangor, C., and Thompson, E. P. (2002). Needs for cognitive economy and self-enhancement as unique predictors of intergroup attitudes. *European Journal of Social Psychology*, *32*, 563–575.

Stiglitz, J. E. (2000). The contribution of the economics of information to twentieth century economics. *Quarterly Journal of Economics*, *115*, 1441–1478.

———. (2002). Information and the change in the paradigm in economics. *Quarterly Journal of Economics*, *92*, 460–501.

Story, L. (2005, September 20). Many women at elite colleges set career path to motherhood, *New York Times*, p. A1.

Sunstein, C. R. (1991). Why markets don't stop discrimination. *Social Philosophy and Policy*, *8*, 22–37.

———. (2006). A new progressivism. *Stanford Law and Policy Review*, *17*, 197–232.

Swim, J. K., and Hyers, L. L. (1999). Excuse me—What did you just say?!: Women's public and private responses to sexist remarks. *Journal of Experimental Social Psychology*, *35*, 68–88.

Swim, J. K., and Sanna, L. J. (1996). He's skilled, she's lucky: A meta-analysis of observers' attributions for women's and men's successes and failures. *Personality and Social Psychology Bulletin*, *22*, 507–519.

Taylor, S. E., and Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, *103*, 193–210.

Taylor, S. E., Fiske, S. T., Etcoff, N. L., and Ruderman, A. (1978). Categorical bases of person memory and stereotyping. *Journal of Personality and Social Psychology*, *36*, 778–793.

Turner, J. C. (1991). *Social influence*. Pacific Grove, CA: Brooks/Cole Publishing.

Uhlmann, E. L., and Cohen, G. L. (2005). Constructed criteria: Redefining merit to justify discrimination. *Psychological Science*, *16*, 474–480.

U.S. Bureau of Labor Statistics. (2006). *2006 annual averages: Household data*. Retrieved from http://www.bls.gov/cps/cps_aa2006.htm

U.S. Equal Employment Opportunity Commission (EEOC). (2006). *Sex-based charges FY 1997–FY 2011*. Retrieved from http://www.eeoc.gov/eeoc/statistics/enforcement/sex.cfm

U.S. General Accounting Office. (2001). *Women in management: Analysis of selected data from the current population survey* (GAO-02-156). Retrieved from http://www.gao.gov/new.items/d02156.pdf

van Knippenberg, A., Dijksterhuis, A., and Vermueulen, D. (1999). Judgment and memory of a criminal act: The effects of stereotypes and cognitive load. *European Journal of Social Psychology*, *29*, 191–201.

Wax, A. L. (1999). Discrimination as accident. *Indiana Law Journal*, *74*, 1129–1231.

Wegner, D. M. (1994). Ironic processes of mental control. *Psychological Review*, *101*, 34–52.

White, R. W. (1959). Motivation reconsidered: The concepts of competence. *Psychological Review*, *66*, 297–333.

Wilson, T. D., and Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations, *Psychological Bulletin*, *116*, 117–142.

Wojciszke, B. (1997). Parallels between competence- versus morality-related traits and individualistic versus collectivistic values. *European Journal of Social Psychology*, *27*, 245–256.

Word, C. O., Zanna, M. P., and Cooper, J. (1974). The nonverbal mediation of self-fulfilling prophecies in interracial interaction. *Journal of Experimental Social Psychology*, *10*, 109–120.

Wyer, N. A., Sherman, J. W., and Stroessner, S. J. (2000). The roles of motivation and ability in controlling the consequences of stereotype suppression. *Personality and Social Psychology Bulletin*, *26*, 13–25.

CHAPTER 4

The Psychology of Cooperation

Implications for Public Policy

TOM TYLER

Across the social sciences there has been widespread recognition that it is important to understand how to motivate cooperation on the part of the people within group settings (Tyler, in press-b). This is the case irrespective of whether those settings are small groups, organizations, communities, or societies. Studies in management show that work organizations benefit when their members actively work for company success. Within law, research shows that crime and problems of community disorder are difficult to solve without the active involvement of community residents. Political scientists recognize the importance of public involvement in building both viable communities and strong societies. And those in public policy have identified the value of cooperation in the process of policy making, such as in stakeholder policy-making groups. For example, efforts to solve long-term environmental problems ranging from the location of nuclear waste dumps to preventing global warming focus upon how to obtain cooperation from different individuals and groups (Weber, this volume).

Cooperation

One obvious way that groups, organizations, and societies want their members to cooperate is by adhering to group rules. This idea is so central to the area of regulation that compliance is viewed as a litmus test for the effectiveness of regulatory authorities. A second important aspect of cooperation is working with the group to achieve group goals. This involves productivity, that is, behavior that helps the group by producing desirable resources.

This distinction between two functions of cooperative behaviors differentiates between those that proactively advance the group's goals through performance of actions that help the group (productivity) and those that limit behaviors that are obstacles to achieving the group's goals (regulation). What this suggests is that the people in groups can cooperate with groups both by doing things that help the group in a positive, proactive, manner and also by refraining from doing things that hurt the group. An employee on the job helps their company both by doing their work well and by not stealing office supplies.

Beyond Material Self-Interest

As already noted, psychologists distinguish between motivations, which are the goals that energize and direct behavior, and people's judgments about the nature of the world. These judgments tell people how to make plans and choose when to take action and how to behave, that is, which decisions or choices to make. Goals are the objectives that motivate behavior—the end states that people value and seek to obtain.

The issue of cognition, or judgment and decision making, involves decisions that people must make about how to most effectively achieve desired goals. It explores how, once they have a goal, people make decisions about how and when to act so as to be most likely to achieve their goals. Motivation explores the issue of what goals people desire to achieve. People's goals, that is, the things that they seek to obtain, help us to understand what motivates their behavior. Unless we know what goal people are pursuing, we cannot understand the intention of their actions. Of course, people may make errors that lead them to fail to achieve their desired goals. Nonetheless, their actions are guided by their purposes.

A simple example of the distinction between cognition and motivation is found in the instrumental analyses of action. The goal that energizes people within instrumental models is the desire to maximize their material rewards and minimize their material costs, such as the punishments that they experience. To do so, people make estimates of the likely gains and losses associated with different types of actions. These judgments about the nature of the world shape the degree to which people engage in different behaviors

in pursuit of their goal of maximizing rewards and minimizing punishments.

Within the arenas of law, management, political science, and public policy, most discussions of human motivation are drawn from the fields of psychology and economics. The assumption that people are seeking to maximize their personal material utilities underlies much of the recent theory and research in both psychology and economics. The argument is that people are motivated by this desire but simplify their calculations when seeking to maximize their personal utilities by satisficing and using heuristics. Furthermore, while motivated to maximize their utilities, people also have limits as information processors, leading them to make errors and act on biases. In other words, people may be trying to calculate their self-interest in optimal ways but lack the ability to do so well. So, people are acting out of the desire to maximize their own material self-interest, but they do it imperfectly due to limits in their time, information, and cognitive abilities.

The Interface of Psychology and Economics

In the past several decades there have been tremendous advances in the connection between economics and psychology. Economists have drawn upon the research and insights of psychologists and have also conducted their own empirical research as part of the burgeoning field of behavioral economics. The goal of this chapter is to further the connection between psychology and economics by showing the value of broadening the range of motivations that are important in social settings.

A major area of psychology upon which economists have drawn during the last several decades is judgment and decision making. This area, characterized by the work of psychologists such as Tversky and Kahneman (1974, 1981; also see Kahneman, Slovic, and Tversky, 1982; Kahneman and Tversky, 2000), focuses upon the cognitive errors and biases that shape the individual judgments of people seeking to pursue their material self-interest during decision making (Brocas and Carrillo, 2003; Dawes, 1988; Hastie and Dawes, 2001; Hogarth, 1980; Nisbett and Ross, 1980; Plous, 1993; Thaler, 1991).

The literature on judgment and decision making is not primarily focused on issues of motivation, but rather on those of cognition. It seeks to understand how people use their subjective knowledge of the world to make decisions. It assumes that a key motivation for such actions is the desire to maximize material gains and minimize material losses. However, an important message from social psychology is that both cognition and motivation are important. They act in tandem to shape behaviors such as cooperation (see Higgins and Kruglanski, 2001). As a consequence, this analysis of decision making can profit from being combined with an expanded analysis of the possible motivations for action.

In terms of motivations, I have noted that economists have generally operated on the assumption that people are motivated to maximize their own personal self-interest, a self-interest defined in terms of material gains and losses. No doubt most psychologists and economists would acknowledge that people can be motivated by a broader range of motivations than material gains and losses, but these other motivations have not been the primary focus of this research. Similarly, the models of human motivation dominating law, management, political science, and public policy have not generally been broader in their focus than to consider the role of incentives and sanctions (Green and Shapiro, 1994; Pfeffer, 1994).

While incentives and sanctions have dominated the study of motivation (Goodin, 1996), there have been suggestions of the need for a broader focus. In articulating such a broader vision of human motivation, this work connects to the recent work of behavioral economists working in this area (among others Akerlof, 2006; Falk and Kosfeld, 2004; Fehr and Falk, 2002; Fehr and Gächter, 2002; Fehr and Rockenbach, 2003; Frey 1997; Frey and Stutzer, 2002; Stutzer and Lalive, 2004). It does so by arguing for the potential benefits of those involved in studying law, management, political science, and public policy of considering a broader range of the motivations that can shape behavior in institutional settings.

Much of this research has been experimental and has focused on demonstrating that social motivations matter. However, such methods are not ideal for evaluating the relative impact of instrumental and social motivations in real-world settings, since the strength of the influence of particular factors in experiments is not necessarily a reflection of the strength of their influence in real-world settings. Rather, in experiments the strength of influence is linked to the strength of the manipulation. The studies I will discuss are surveys that explore the natural range of instrumental and social factors, as well as their relative influence in social settings. I will argue that these methods more effectively demonstrate that social motivations not only matter but also have a strong influence upon behavior in social settings.

A New Framework for Voluntary Cooperation

The aim of this chapter is to build upon the suggestion that we need a better understanding of the

variety of factors that shape people's behavior in social settings. In particular, we need to provide a better sense of what social motivations are and how they influence behavior. To do so, I will present an analysis of several studies and, through them, move toward a broader framework for understanding human motivation within social settings. That framework includes concern with costs and benefits and with issues such as reputation as defined by economists.

As framed in traditional economic analysis, these issues are instrumental and are linked to concerns about the material gains and losses that are expected to result from different possible behaviors. The studies reviewed here suggest that social mechanisms help us to explain additional aspects of voluntary cooperation beyond the pursuit of material gains and the avoidance of material losses. These social mechanisms provide another set of goals or reasons that lead people to take the actions that they do when they are in social settings.

In this analysis I will identify several types of social psychological mechanisms that deal with issues relevant to cooperation and test their importance in a community and an organizational setting. This demonstration is based upon the premise that it is through showing that social motivations influence cooperation that their value in social settings can best be demonstrated. The core argument is that although people are clearly motivated by self-interest, and seek to maximize their material rewards and minimize their material deprivations, there is a rich set of other, more social motivations that also shape people's actions. These motivations have an important influence on behavior that is distinct from instrumental calculations but have not received as much attention as have material gains and losses. The argument that social motives are important has implications for issues of the design of groups, organizations, and communities. The primary implication is that there are a broader range of motivations that can be tapped to encourage desirable behaviors than is encompassed within traditional incentive and sanctioning models.

Instrumental Models

The traditional focus of motivational discussions is the instrumental model, which posits that people engage in interactions to exchange material resources. That is, behavior is motivated by offering incentives for desired behavior, threatening sanctions for undesired behavior, or both. There are a variety of such models. One is the rational choice approach, in which people are viewed as motivated by incentives and sanctions (for an extended treatment of this approach see Darley and Alter, this volume). A more complex version of this approach is based upon resources invested in attaining one's position in a group, which is linked to the expectation of long-term gains. A third type is an instrumental approach that focuses on the dependence that people have on the resources they receive from their job. A fourth model consists of instrumental models of justice, which are concerned with using distributive justice principles to maximize long-term outcomes when dealing with others. Finally, there is instrumental trust, which is concerned with expectations about the likely behavior of others.

Social Motivations

A contrasting type of motivation is social. This section will identify, measure, and show the importance of five types of social motivations. Those social motivations are attitudes, values, identity, procedural justice, and motive-based trust. In reviewing these social motivations, the general goal is to demonstrate the benefits of moving beyond using only material self-interest to motivate cooperation.

Attitudes: Commitment/Intrinsic Motivation

Attitudes are internal predispositions—part belief, part feeling—that motivate people to engage in actions that are personally fulfilling by approaching those objects, activities, or events toward which they have favorable beliefs and positive feelings, and to avoid those about which their orientation is negative (Brief, 1998). Attitudes reflect the behaviors that people want to engage in, whereas values reflect those actions that they feel they ought to engage in. Both attitudes and values are long-term predispositions that shape behavior in a manner that is distinct from the influence of the benefits and costs associated with immediate environments. The concept of attitudes has a long history within social psychology (McGuire, 1969). As would be expected given both that long history and the breadth of the study of attitudes, there are many variations in how attitudes are defined and studied. Here attitudes will be thought of as general predispositions for or against a group or activity, predispositions that are acquired over time and expressed across situations (Brief, 1998).

Social Values

A second body of literature concerns social values. Values reflect people's feelings of obligation and responsibility to others. One example of an obligation-based judgment is the assessment of the legitimacy

of a set of rules or authorities. Legitimacy has been shown to be a predictor of rule-following behavior in both communities (Tyler, 2006b; Tyler and Huo, 2002) and work organizations (Tyler, 2005; Tyler and Blader, 2005; Tyler, Callahan, and Frost, 2007). Obligation is often studied in the context of rule following, but people can also be motivated by their feelings of obligation to perform well on the job, as well as to be loyal to their firm.

A second form of obligation is the people's feelings of obligation to their own moral values—their desire to do what they feel is right. A large psychological literature argues that the motivation to act in accord with their values about what is morally right and wrong is an important human motivation (see Darley and Alter, this volume; Tyler and Darley, 2000).

Identity

Another arena within which instrumental and social motivations can be contrasted is the general relationship between people and organizations. Social psychology has long asked the question, What links people to groups, organizations, and communities? This question can be answered by reference to people's desire to create and maintain a favorable identity and a positive sense of self. Theories linked to identity argue that organizations serve the social function of providing people with an identity, which they draw at least in part from the organizations to which they belong (Tyler and Blader, 2000). That argument can be made based upon group-based social identity or emotional identification. Emotional identifications are links to other people whose nature shapes people's identities.

The argument that groups serve a role in defining and evaluating one's identity, and thereby shaping one's sense of self, is a key suggestion of models of social identification (Hogg and Abrams, 1988). A second group of identity-based literature concerns emotional identification. Since the pioneering work of Kelman on persuasion and attitude change (Kelman, 1958, 1961), it has been recognized that emotional ties to others provide an important input into identity (Kelman, 2006; Kelman and Hamilton, 1989).

Procedural Justice

An additional social motivation can be drawn from the literature on justice (Tyler, 2000). There are two key forms of justice: distributive and procedural. Distributive justice refers to the fairness of outcomes, while procedural justice reflects the justice of the procedures through which outcomes are distributed and decisions made.[1] The justice literature suggests that people's procedural justice judgments are relational (i.e., social) in nature and are linked to their social connections to others (Tyler, 1994, 2000; Tyler, Degoey, and Smith, 1996). Hence, the influence of procedural justice upon cooperation is an instance of the impact of social motivation.

Motive-Based Trust

Finally, the literature on trust suggests that people are more willing to cooperate with others whom they trust (Kramer, 1999; Kramer and Tyler, 1996; Tyler and Huo, 2002). One type of trust, motive-based trust, is linked to inferences about the character of the other. This form of trust is also linked to relationships to others and is a social motive.[2]

What Leads a Motivation to Be a Social Motivation?

This chapter characterizes the additional motivations described—attitudes, values, identity, procedural justice, and motive-based trust—as social motivations, to distinguish them from instrumental motivations. As has been noted, instrumental motivations reflect people's desire to gain material resources and to avoid material losses. Social motives, as discussed by psychologists, differ in that they are motivations that flow from within the person.

There are four ways to distinguish between instrumental and social motivations. The first is by the content of the concerns that people express within each domain. Instrumental concerns focus on the potential for material gains and the possibility of material losses. Such gains and losses involve gains in terms of rewards and losses in terms of costs or punishments. In contrast, social motivations are linked to gains and losses of a nonmaterial nature. Such gains and losses are linked to such issues as personal identity and consistency with ethical and moral values.

Second, indicators of social motivations are empirically distinct from indicators of material gain or loss. For example, in the literature on social justice, it has been found that people distinguish between receiving a favorable outcome and receiving fair treatment (Tyler et al., 1996). Hence, judgments about justice are distinct from the favorability of one's outcomes. This distinction is clear in the literature on distributive justice, a literature in which the fairness of outcomes is distinguished from their desirability (Walster, Walster, and Berscheid, 1978). It is even clearer in the

literature on procedural justice, which focuses on the fairness of the procedures by which allocation decisions are made (Lind and Tyler, 1988). If people simply viewed a favorable outcome as fair, for example, social motivations would not be distinct from material judgments. However, this is not the case.

Third, social motivations have a distinct influence on cooperative behavior. Again, the justice literature finds that the degree to which people are willing to accept an outcome from an authority is linked, first, to the favorability of that outcome. In addition, however, people are more willing to accept an outcome that they evaluate as being fair and fairly arrived at. Hence, the outcome fairness of judgments exercises an independent influence upon outcome acceptance behavior that cannot be explained by outcome favorability. Similarly, procedural fairness has a distinct influence on acceptance behavior.

Fourth, social motivations produce consistent behavior across situations and over time. If, for example, someone feels an obligation to obey rules, their behavior should consistently reflect higher levels of cooperation across settings that vary in their reward and cost characteristics. Furthermore, they should show the same consistency of behavior in the same situation across time. This does not mean that situational forces will not influence behavior, but it will be possible to see constancies in behavior that are not linked to those forces.

The best way to understand the value of this larger motivational framework is to consider it within the context of a particular type of socially important behavior, so I will focus on the motivation of cooperative behavior. Cooperation is valuable for groups, and securing cooperation has been the focus of social science research across a variety of fields, including both economics and psychology (see Tyler, in press-b, for a review). Hence, the question of how to motivate people to cooperate is a key concern within public policy.

Why Does the Strategy of Motivation Matter?

While a discussion of the factors that motivate people can seem abstract, the way this question is answered has important public policy implications. Consider the arena of regulation. The United States is currently committed to a costly strategy of trying to reduce crime by deploying legal authorities to create a credible risk of being caught and punished for rule breaking, coupled with a large prison system to punish those who are caught. In fact, our society is a leader in the world in the proportion of its population that is in prison, creating a massive expenditure of resources. Yet, this costly strategy is based upon very little compelling evidence either that it works or that it is the best possible strategy. So, a wide variety of social policies, ranging from policing strategy to sentencing practices and correctional and rehabilitative models are linked to underlying assumptions about what works in the area of motivation. And, as I will show below, considerable evidence exists to suggest that the current motivational models are badly off track (Tyler and Rankin, in press-a).

Although regulation is the most striking example of the issue of model misspecification, even in work organizations in which people can be both incentivized and sanctioned, rewards and punishments are not found to be a particularly effective motivator of work performance, particularly when performance involves creative or innovative problem-solving behavior. Incentives work best for routinized tasks, whereas work environments in the United States are increasingly white-collar jobs requiring creativity and innovation. Hence, in the broad arena of work organizations there is a similar misspecification of motivational models. This has widespread policy implications. For example, the reaction to corporate scandals has been to focus upon shifting incentives rather than upon building values of honesty and integrity. This instinctive policy resort to instrumental mechanisms seems widespread within the business world, but I will argue it is oversimplistic at best and misguided at worst. If the focus were shifted to social motivations, we would become concerned with building strong organizational cultures that emphasize values and identity. In other words, as with regulation, many of our reactions to public policy issues are shaped by our models of motivations, models that are often based more strongly upon assumptions than upon empirical facts.

Comparing Motivational Strategies Empirically

Regulatory Settings

The first setting in which I consider the antecedents of cooperation is a community setting in which authorities are seeking help from community members. In this study the authorities involved are the police and the help being sought is voluntary cooperation of two types: support for police efforts to deal with crime and a willingness to work with the police to manage social order in the community. I refer to this setting as regulatory because it focuses upon efforts to enforce legal regulations against criminal behavior. However, my concern here goes beyond simply securing compliance with the law. I am concerned with

gaining the active voluntary cooperation of community residents to fight crime.

DESIGN

The study in question was a panel study in which New Yorkers were interviewed over the telephone at two different times (for details see Sunshine and Tyler, 2003; Tyler and Fagan, 2008). The time-one sample consisted of 1,653 interviews with residents of New York City. The sample was drawn from a stratified random sample of telephone numbers in the city, with an overrepresentation of non-White residents designed to produce a high proportion of Hispanic and African American respondents. Approximately one year following the first interview, attempts were made to recontact and reinterview all of the respondents. A subset of 830 of those originally interviewed was successfully reinterviewed. A comparison of those reinterviewed with the original sample indicated no significant differences in ethnicity, gender, age, income, or education. Both the sampling and questionnaire details can be found in Tyler (in press-b).

COOPERATION

Cooperation involved voluntary efforts to help the police. It was assessed by asking respondents whether, if the situation arose, they would be likely to call the police to report a crime; help the police to find someone suspected of a crime; report dangerous or suspicious activity; volunteer time to help the police; patrol the streets as part of an organized group; or volunteer to attend community meetings to discuss crime.

SOCIAL MOTIVATIONS

Five social motivations were measured: attitudes about the law; values (i.e., the legitimacy of the police and law and the congruence of law with moral values); identification with the police; the procedural justice of the behavior of the police (overall quality of decision making and quality of treatment); and motive-based trust.

ANALYSIS

Instrumental and social judgments were found to be related ($r = 0.26$). Those members of the community who believed that the police were instrumentally effective also expressed greater social motivations for cooperating with them.

Structural equation modeling was used to examine the influence of instrumental and social factors in shaping cooperation. That analysis included separate analyses for the wave-one sample and for the panel sample. Because of the community context and the ethnic diversity of the sample, demographic variables were included in this analysis.[3]

Table 4.1 Motivation in a community setting

	Voluntary cooperation	
	Time one	Time two
Time one factors		
Demographics	0.23***	0.06
Instrumental motivations	0.17***	0.03
Social motivations	0.25***	0.11**
Cooperation	—	0.74***
Adjusted explained variance	14%	58%
	1,653	830

The results for the wave-one cross-sectional analysis, shown in Table 4.1, suggest that both instrumental and social factors play an important role in shaping cooperation. There were strong influences of social motivations (beta = 0.25, $p < .001$); demographics (beta = 0.17, $p < .001$); and instrumental motivations (beta = 0.17, $p < .001$). Interestingly, when controls were placed upon prior cooperation in a panel analysis, only social motivations continued to show an independent influence (beta = 0.11, $p < .01$). The panel results suggest that social motivation remains important with a more sensitive panel design, while the influence of instrumental variables disappears.

POLICY IMPLICATIONS

The results of this study reveal that there is a generally high level of willingness to cooperate with the police. For example, the mean willingness to report crimes to the police in the panel sample was 3.57 on a scale of 1 (not likely at all) to 4 (very likely), while the mean for working with people in the community was 2.79 on the same scale (see Tyler and Fagan, 2008). In other words, there was generally considerable willingness to help the police, something that can be built upon in efforts to manage crime and social order in both majority and minority communities.

The issue is how to engage this potential cooperation. A report from the National Academy of Sciences noted the paradox that while the police had improved their objective performance in recent decades and crime rates had declined, public trust and confidence in the police had not improved appreciably, especially in the minority community (Skogan and Frydl, 2004). The report recommended a focus on social motivations that create or undermine trust and argued that attention to social motivations would

enhance both trust and public cooperation. That argument is supported by a number of studies of policing (for an overview, see Tyler, in press-b).

Factors Shaping Cooperation in Work Settings

The second setting I will consider involves employees in for-profit work settings. Again, a key issue is cooperation within the organization.

COOPERATION IN WORK SETTINGS

Four forms of cooperation were distinguished: in-role behavior—that is, doing one's specified job; extra-role behavior—that is, engaging in behavior to help the group beyond what is required; voluntary compliance behavior—that is, adhering to rules; and following rules.

INSTRUMENTAL MOTIVATIONS

Five forms of instrumental motivation were measured.

Environmental contingencies. This form of instrumental motivation was assessed in two ways. First, respondents were asked about the strength of the connection between good/bad workplace behavior and incentives/sanctions—that is, the likelihood that good performance would be rewarded and rule breaking punished. Second, they were asked about the magnitude of the incentives/sanctions linked to good/bad behavior. Finally, the interaction between these two judgments was measured.

Investment. The long-term possibilities for gain through the company and the favorability of company policies were measured.

Dependence. People were asked whether their orientation toward work was instrumental—that is, they worked only for money—and whether they needed their job for financial reasons.

Distributive fairness. Distributive fairness was assessed at two levels: organizational and personal. These judgments reflect the degree to which employees felt rewards and opportunities were distributed fairly in their organization.

Instrumental trust. Calculative trust was measured. Calculative trust is an estimate by the respondent of the likelihood that others will be trustworthy if they are trusted.

SOCIAL MOTIVATIONS

Five types of social motivations were considered.

Attitudes. Three aspects of attitudes were considered: attitudes toward the company; attitudes toward one's job; and work-related emotion/affect.

Values. Feelings of obligation were measured in four ways: the legitimacy of workplace rules; the degree to which the respondent felt obligated to deliver high-performance work; the degree to which the respondent felt obligated to stay at their current work organization; and the congruence of company policies with the employee's moral values.

Identity indicators. This analysis drew upon the conceptualization of Tyler (Tyler and Blader, 2000; Tyler and Smith, 1999) and operationalized identity in terms of respect, pride, and identification. In the case of pride, the measurement looked at pride linked to the status of the group (Tyler and Blader, 2000), while respect referred to status in the organization, and identification to the degree that an employee merged their sense of self with their company. In addition emotional identification, as conceptualized by Kelman (1958), was measured.

Procedural justice. Procedural justice was measured in four ways: the general procedural justice of the organization; the personal procedural justice experienced at the organization; supervisor procedural justice of decision making/interpersonal treatment; and organizational-level procedural justice of decision making/interpersonal treatment.

Motive-based trust and cooperation. In this analysis three indices were used to measure motive based trust. Those indices measured trust in the motivations of organizational authorities; overall trust in management; and trust in one's supervisor.

ANALYSIS

Because instrumental and social motivations were linked to each other, it was important to consider the simultaneous independent contribution of each type of motivation to cooperate. Regression analysis was conducted within the framework of structural equation modeling, which allowed the indices of the five instrumental and the five social clusters to load upon two latent factors: instrumental and social. Those underlying, or latent, factors were then used to predict cooperation. Cooperation was also an underlying, or latent, factor, reflecting the four indices of cooperation. As with the community sample, this analysis was first conducted on the wave-one respondents and then on the panel respondents. In the second analysis, cooperation in wave one was controlled upon when explaining cooperation in wave two. The results of these analyses are shown in Table 4.2.

In the time-one analysis, the estimated influence of social motivations was beta = 0.62, $p < .001$; and for instrumental factors beta = 0.04, $p < .001$. When voluntary cooperation was distinguished from required cooperation, the influence of social motivations was

Table 4.2 The influence of instrumental and social motivation on cooperation

Time one cooperation (4,430)			
	Total	Required	Voluntary
Time-one factors			
Instrumental: time one	.04***	0.11***	0.16***
Social: time one	.62***	0.53***	0.75***
Adj. R-sq.	38%	29%	59%
Time two cooperation (2,680)			
	Total	Required	Voluntary
Time-one factors			
Instrumental: time one	0.01	0.03	-.01
Social: time one	.11***	0.11***	0.23***
Cooperation: time one	.76***	0.65***	0.69***
Adj. R-sq.	59%	45%	53%

found to be stronger when voluntary cooperation was being examined. An analysis of the panel data indicates that social motivations continued to be important, while instrumental influences became relatively less important.

POLICY IMPLICATIONS

These findings suggest that we can more effectively explain cooperation by considering social and instrumental motivations than by focusing solely on instrumental motivations. Social motivations are always found to be influential, irrespective of whether cross-sectional or panel analyses are considered. In fact, social motivations are the dominant factor shaping cooperation.

On average, the score for required cooperation was 6.19, while the average for voluntary cooperation was 4.78 (on a scale ranging from 1 to 7). This distinction was greater with performance. The mean for in-role performance was 6.46, and for extra-role behavior, 5.46, a difference of 1.00. In the case of rule following, the difference between required and voluntary levels of cooperation was 0.45. In general, these means showed that respondents indicated that they engaged in generally high levels of cooperation. As would be expected, voluntary forms of cooperation were less common than were required forms. In other words, as was true with community residents and the police, employees show considerable willingness to cooperate with managers, even when we consider voluntary forms of cooperation. And this willingness is enhanced by social factors, such as fair workplace procedures and trustworthy managers.

Conclusions and Implications

This chapter focuses on a micro-level exploration of the behavior of the people within organizations. As noted, such an approach is premised upon the belief that the behavior of the people within groups shapes the viability of those groups. The suggestion that the thoughts, feelings, and behavior of the people within groups are linked with group viability and functioning is widely supported by studies within law (Tyler, 2006b), management (Allen and Rush, 1998; Freund and Epstein, 1984; Katz, Kochan, and Weber, 1985; Koys, 2001; Pfeffer, 1994; Podsakoff, Ahearne, and MacKenzie, 1997; Shore, Barksdale, and Shore, 1995), and public policy and government (Culpepper, 2003; Harrison and Huntington, 2000). In each area it has been shown that the beliefs, feelings, and behaviors of group members influence the functioning of groups.

The goal of organizational design is to produce groups, organizations, and communities that are effective, efficient, and satisfying to their members. When these goals are not being achieved, one clear implication is that the structure of the group should be altered. This premise, that the structure of a group, organization, or community shapes the behavior of the people within it, and through that, influences the group, is a core assumption of social psychology, which both views human behavior as a response to the nature of the social institutions within which people are embedded and suggests that the viability of groups is linked to the behavior of the people within them.

A consequence of this argument is the suggestion that when the design of organizations is not consistent with the realities of human motivation, organizations have difficulty achieving their objectives (Ferraro, Pfeffer, and Sutton, 2005). As Weber argues (this volume), the mindset of our society, that is, its assumptions about people's psychology, stacks the cards in a variety of ways against effectively achieving cooperation, including the poor framing of instrumentally motivated choices, which interferes with the ability to be rationally self-interested, and the neglect of issues of responsibility, obligation, and moral values, which minimizes the influence of social motivations on cooperation.

Tyler (2007) suggests that the legal system has difficulty effectively motivating rule adherence because its model of human motivation does not encompass the primary factors that actually shape deference to rules. Research suggests that people's rule-related

behavior is most strongly influenced by their sense of responsibility and obligation to defer to legitimate authorities and follow moral principles (Tyler, 2006b; Tyler and Blader, 2005; Tyler, Callahan, and Frost, 2007). However, legal institutions are designed on the assumption that behavior is shaped by the instrumental risk of sanctioning. As a result, there is a fundamental misalignment of the organization, in this case the legal system, and models of motivation, leading the system to be less efficient and effective than might potentially be the case.

Similarly, in the case of the management of work organizations, there is a strong emphasis upon the use of incentives and sanctions to motivate desired workplace behavior. Although studies suggest that the combination of incentives and sanctions is better able to motivate behavior than is a focus only on sanctions (Podsakoff et al., 2005; Tyler and Blader, 2000), the ability of instrumental variations to shape cooperation is still limited. Again, command-and-control approaches reflect a misalignment of institutional design with human motivation. This misalignment occurs not because instrumental models cannot predict factors that influence cooperation, but because a focus only on instrumental issues is incomplete.

This chapter provides the authorities in groups, organizations, and communities with a perspective on how to better motivate desirable behavior on the part of the members of the groups they manage. The exercise of authority can potentially involve many tasks, and this discussion does not deal with all of them. It focuses, instead, on one concern that is common to many group situations: shaping the cooperative behavior of the people within a group, organization, or community. This chapter has been concerned with factors shaping people's motivation within groups, organizations, and communities.

I do not argue with the suggestion of instrumental models that incentives and sanctions can and often do shape cooperation. Such effects are widely, but not universally, found. Rather, I argue that an exclusive focus on instrumental approaches is not optimal, since social motivations are the strongest drivers of cooperative behavior. Hence, organizations that rely primarily or exclusively upon instrumental motivations are misaligned with human motivations and not optimally designed.

Instrumental models have the advantage of being under the control of authorities, as long as those authorities can obtain and maintain the wherewithal to deploy them. Hence, they are a motivational "sure thing," as least while times in a group are good. It may be this element of control that most strongly draws authorities to instrumental approaches. And with the control of resources comes organizational centrality, since when authorities use instrumental approaches, they become the focus of attention, with employees shaping their behavior in response to what authorities are doing rather than in response to their own needs and concerns. Because instrumental approaches easily resonate with authorities, one approach that might be taken by them to try motivating cooperation would be to build up the effectiveness of instrumental approaches to maximize their ability to motivate members of communities and organizations.

However, the approach to increasing the motivation to cooperate I have taken in this chapter is not to strengthen instrumental approaches but instead to broaden the conception of what motivates employees. By including social motivations in the overall motivational framework, the ability to explain why people cooperate is substantially enhanced. The implication for organizational design is that there needs to be a focus on creating the organizational conditions conducive to promoting social motivations.

The argument outlined here is based upon the distinction in utility functions between the strategies that people use to achieve their objectives and the nature of their objectives—the end states that people value. The judgment and decision-making literature has made clear in the last several decades that there is a great deal to be gained by exploring the individual's thought processes—by developing the expectancy aspect of utility. This chapter has suggested that there is a similar benefit to developing the second aspect of the utility model—our understanding of what people value, that is, creating an expanded version of the goals that motivate people in social settings. While they are motivated by material incentives, such as opportunities for pay and promotion, and seek to avoid losses, such as sanctions for rule breaking, people are motivated by a broader set of issues, organized here and labeled social motivations. Those social motivations include attitudes, values, emotions, identity, procedural justice, and motive-based trust.

Implications for Organizational Design

The empirical analysis makes clear that instrumental and social motivations are related but that each has distinct influences upon cooperation. Hence, neither is simply a reflection of the other. It is also the case that each form of motivation has distinct conceptual characteristics and that, as a result, they are not interchangeable from an organizational-design perspective.

Instrumental approaches are widely used because they have the advantage of providing a reliable approach to motivation. Authorities do not need to seek to understand the people over whom they are

exercising authority. They can deploy a system of incentives and sanctions that will generally shape behavior in the directions that they desire.

Discussions about the use of incentives and sanctions consistently point to several limitations of such motivational approaches. One that has already been noted is that their impact is weak. When they work, these strategies have modest effects upon behavior.

Second, even when they are being successful, their use requires the continual deployment of organizational resources. Employees whose work is motivated by pay do not over time become motivated by other factors. On the contrary, their internal motivations to work are undermined by an emphasis upon pay for performance. Hence, the organization must continually allocate resources to maintain performance and may, over time, gradually have to increase those allocations as other reasons for working are undermined.

The case of resource drain is especially striking with regulation. To prevent wrongdoing, organizations need to deploy a credible surveillance effort, such as a police or security force. And since the likelihood of apprehension is the primary determinant of behavior, that force needs to be of sufficient size and quality that it presents people with a reasonably high level of risk of being caught and punished for wrongdoing. Preventing rule breaking is essential for viability, but it does not directly add value to the organization. Furthermore, it is often out of the control of group authorities.

The September 11 attack upon the World Trade Center forced a massive reallocation of resources to defense and policing within the United States at a time when authorities would have preferred to deploy those resources in ways that would be of greater benefit to our society—to improve health care, lower the deficit, etc. Hence, resources deployed for regulation cannot be optimally used to promote effectiveness and are deployed reactively and out of necessity to combat security threats. While communities may gain economically by having a prison constructed and guards employed in their community, spending money to prevent and punish rule breaking is not the optimal use of collective resources. Having a large standing army or police force may be necessary, but it is costly and, over time, drains resources from an organization. Similarly, the resources that work organizations spend to monitor employees for theft and other types of rule breaking are necessary to counter these serious organizational problems but are not a desirable use of resources.

A further problem with instrumental approaches is that they are least effective when they are needed most. All groups have periods of difficulty and scarcity. A company may have a decline in market share and may need to reorganize and rebuild. A community may suffer drought or flooding and require sacrifices from its members to remain viable. It is at these times that the cooperation of group members is most essential to the survival of the group. But, ironically, this is when that cooperation is least likely to be obtainable via instrumental approaches.

A study by Brann and Foddy (1988) is illustrative. In their study, members of a community were told that collective resources (fish in the pond) were being depleted too rapidly and might disappear. Those motivated by self-interest reacted to this information by increasing the number of fish they took from the pool. Those socially motivated took fewer fish. This study illustrates how instrumental motivations may often lead people to act against the group when the group is vulnerable rather than sacrificing on behalf of the group at the risk of their self-interest. From the perspective of the group, the viability of the group is more uncertain when it contains within it people primarily motivated by self-interest.

Tyler (2006a) reviewed the literature on legitimacy and found that, as this argument would suggest, groups whose leaders are legitimate and who therefore have a basis upon which to call upon their members to make sacrifices are more likely to survive periods of difficulty. Those leaders can appeal to social motivations when they lack the resources to reward sacrifice or punish rule-breaking behavior. So, they have to use alternative approaches to motivate cooperation during troubled times.

Social motivations are conceptually distinct from instrumental motivations, and as a consequence they have distinct strengths and weaknesses. A distinct strength is, as has been noted, that they do not require organizational authorities to possess the ability to provide incentives for desired behavior or to be able to create and maintain a credible system of sanctions. At all times, groups benefit from having more resources available that can be directed toward long-term group goals. If everyday group actions are shaped by self-regulating motivations, groups have more discretionary resources.

And as the findings of this chapter make clear, social motivations are important because they are more powerful and more likely to produce changes in cooperative behavior than are instrumental motivations. Hence, social motivations are both more powerful and less costly than are incentives and sanctions. Of course, this does not mean that social motivations can be immediately and automatically deployed in all situations.

A weakness of social motivations is that they cannot be quickly activated within any social context. A CEO with a million-dollar war chest can create

an incentive system to motivate behavior in desired directions overnight. Conversely, a city can shift its police patrols around to vary the nature of the threat faced by community residents. Such flexibility is a major advantage of the instrumental system. Social motivations must be developed over time, as the culture of an organization is created. Hence, a long-term strategy is needed to build an organization based upon social motivations.

More simply put, in order to be able to call upon people's loyalty and patriotism when sacrifices are required, it is necessary for loyalty and patriotism to exist widely among members of the group, organization, or society. This requires a long-term strategy to inculcate and maintain such values. The findings of the research outlined here indicate that one element of such a strategy needs to involve efforts by authorities to make decisions in ways that are viewed as procedurally just and that lead to trust in the motivations of those authorities.

A strategy based upon social motivation also has the disadvantage of taking control away from those at the top of the social hierarchy. If a group relies on voluntary cooperation, its leaders need to focus upon the attitudes and values of the people in the group. For example, they have to create work that people experience as exciting. Furthermore, they have to pursue policies that accord with the employees' moral values. These aspects of social motivation create constraints upon the actions of leaders.

It is natural that leaders would prefer a strategy in which they are the focus of attention, irrespective of its effectiveness, to one in which they focus their attention upon the concerns of their employees or constituents. Yet, within business organizations, a focus on the customer is a widely institutionalized value. Similarly, the concept underlying democratic processes is that, within communities, policies ought to be a reflection of the values of the members of those communities. Hence, it is hardly a radical suggestion that organizations benefit when they develop their policies and practices in consultative ways that involve all of the relevant "stakeholders," including leaders, group members, and external clients such as customers.

Aspects of procedural justice and motive-based trust feed directly into the need to make group policies and practices consistent with the attitudes and values of group members. Participatory decision making and consultation at all levels (i.e., opportunities for voice) are mechanisms through which people's views are represented. And one key element in trust is the belief that authorities are soliciting and considering people's views before making decisions. Procedures that are viewed as procedurally just and authorities judged to be trustworthy encourage input from employees to higher management. And, of course, neutrality (making factually based and impartial decisions that consistency apply rules across people) and quality of treatment (respect for people and their rights; treatment with courtesy and dignity) are also important elements of procedural fairness, as well as processes that engender trust in authorities and institutions (for a more detailed discussion of elements of justice and trust, see Tyler, 2007).

Ironically, those constraints may often have additional value for groups. The era of corporate excess makes clear that the unchecked power of those in high management does not always end up serving the interests of the company. Hence, the need to be accountable to others within the organization may have valuable benefits for the group and may check the tendency of leaders to engage in unwise actions. Just as "checks and balances" is frequently held up as one of the primary desirable design features of American government (Tyler, 2007), the balancing of policies and practices among stakeholders has the benefit of restricting any tendency toward excesses.

Building Identification, Creating Attitudes and Values

Social motivations are important in theoretical terms because they point a direction for future research into cooperation. In addition to exploring how to motivate cooperation via more effective systems of sanctioning or by innovative incentive-based strategies, the findings outlined above argue that we would benefit from incorporating other motivations into our motivational models. This process leads to a focus on understanding which social motivations shape cooperation.

In my analysis, identification emerges as a key antecedent of cooperation. How do we build identification? The findings outlined suggest that we create organizations in which people experience procedural justice in their dealings with the authorities and institutions that they trust. These two aspects of the employee's experience in their organization are central to their decision to merge their identities with the identity of their group.

Can this argument be extended to attitudes and values? The results of the study of work organizations suggest that the answer is yes. The procedural justice of the organization is linked to the favorability of attitudes ($r = 0.55$); and of value judgments ($r = 0.62$). Similarly, the degree to which employees trust the motives of their managers is linked to the favorability of attitudes ($r = 0.53$) and of value judgments ($r = 0.59$). In other words, if people work in a justly managed social setting, managed by authorities whose motives they trust, their commitment to the

organization and to their jobs is higher and they feel more obligation toward the authority.

Of course it is not necessary to view these findings as only speaking to issues of organizational design. They also have implications for selection. To the degree possible, it makes sense to recruit and seek to retain those group members whose attitudes, values, and identity are already favorable toward the group. While the group climate clearly shapes cooperation, it is not the only potentially relevant factor.

Cooperation in the Lewinian Tradition

Cooperation, as discussed in this chapter, was conceptualized in the tradition of motivational research begun by Lewin (Gold, 1999) and central to the Research Center for Group Dynamics inspired by that research. In Lewin's classic studies, the focus of concern was the behavior of groups. Various types of behavior were considered, including the performance of group tasks and aggression toward others in the group. In the studies leaders sought to encourage/discourage these behaviors using a variety of styles of motivation, including authoritarian and democratic leadership. Lewin focused his own attention upon issues of aggression and scapegoating, as well as on the performance of group tasks, with many studies centered upon the behavior of adolescents. The focus on group performance carried forward as an important aspect of the agenda of the Center for Group Dynamics was inspired by the work of Lewin and his students.

This analysis has been broadly framed using the field theory model in several ways. First, this analysis of people's actions viewed employee behavior as a reflection of two factors: external (instrumental) and internal (social) motivations. Second, the key issue is the mix of these motivations. Finally, this analysis distinguished between those behaviors that are and those that are not voluntary, that is, behaviors that do and do not occur in settings in which behavior is being observed and those who engage in it are aware that incentives and sanctions will be shaped by their actions.

Summary

This chapter has argued for the value of broadening our conceptualization of the goals that people pursue when they are members of groups, organizations, or societies. Beyond their motivation to obtain material resources, which is shaped by the rewards and sanctions risks in their immediate environment, people are also motivated by social considerations. The results outlined here indicate that such social motivations strongly influence people's cooperation with group authorities and with their rules and policies. They are particularly powerful motivators of voluntary cooperation. Since achieving widespread voluntary cooperation has advantages for groups, it is argued that understanding how to develop and sustain social motivations is an important element in organizational design.

Notes

1. Two arguments frame the suggestion that justice is a social motivation. The first is that procedural justice is distinctly relational. Tyler (1994) distinguished procedural and distributive justice, arguing that procedural justice is uniquely framed by "relational motivations." These relational issues include concern about the quality of decision making and the quality of interpersonal treatment (Tyler and Lind, 1992).

The original discussion of relational motivations included trust. In this analysis, trust in the motives of authorities will be treated separately in the chapter on trust. Treatment of this issue has not always been the same. Tyler and Blader (2000) included indices of trust in their index of interpersonal treatment to create two factors: decision making and interpersonal treatment. On the other hand, in their analysis of personal experiences with authorities, Tyler and Huo (2002) treated both general procedural justice judgments and assessments of trust as distinct judgments about the justice of decision making and the justice of interpersonal treatment. This analysis follows the lead of Tyler and Huo (2002) in treating trust as an issue that is distinct from procedural justice.

2. Of course, trust is not completely distinct from procedural justice (see De Cremer and Tyler, 2007).

3. Prior research indicates that there are large differences in cooperation with the police that are linked to demographic factors, in particular race.

References

Akerlof, G. A. (2007). The missing motivation in macroeconomics. *American Economic Review*, *97*, 5–36.

Allen, T. D., and Rush, M. C. (1998). The effects of organizational citizenship behavior on performance judgments. *Journal of Applied Psychology*, *83*, 247–260.

Brann, P., and Foddy, M. (1988). Trust and consumption of a deteriorating common resource. *Journal of Conflict Resolution*, *31*, 615–630.

Brief, A. P. (1998). *Attitudes in and around organizations.* Thousand Oaks, CA: Sage.

Brocas, I., and Carrillo, J. D. (2003). *The psychology of economic decisions.* Oxford: Oxford University Press.

Culpepper, P. D. (2003). *Creating cooperation: How states*

develop human capital in Europe. Ithaca: Cornell University Press.

Dahl, R. A. (2006). *On political equality*. New Haven: Yale.

Dawes, R. M. (1988). *Rational choice in an uncertain world*. San Diego: Harcourt Brace Jovanovich.

De Cremer, D., and Tyler, T. R. (2007). The effects of trust and procedural justice on cooperation. *Journal of Applied Psychology, 92*, 639–649.

Falk, F., and Kosfeld, M. (2004). *Distrust—The hidden cost of control*. IZA Discussion paper 1203. Institute for the Study of Labor (IZA). Bonn, Germany.

———. (2006). The hidden costs of control. *American Economic Review*, *96*(5), 1611–1630.

Fehr, E., and Falk, A. (2002). A psychological foundation of incentives. *European Economic Review*, *46*, 687–724.

Fehr, E., and Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences*, *8*, 185–190.

Fehr, E., and Gachter, S. (2002). *Do incentive contracts undermine voluntary cooperation?* IZA Working paper 1424–0459.

Fehr, E., and Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature*, *422*, 137–140.

Ferraro, F., Pfeffer, J., and Sutton, R. I. (2004). Economics language and assumptions: How theories can become self-fulfilling. *Academy of Management Review*, *30*, 8–24.

Freund, W. C., and Epstein, E. (1984). *People and productivity*. Homewood, IL: Dow Jones/Irwin.

Frey, B. S. (1997). *Not just for the money: An economic theory of personal motivation*. Cheltenham, UK: Edward Elgar.

Frey, B. S., Benz, M., and Stutzer, A. (2004). Introducing procedural utility: Not only what, but also how matters. *Journal of Institutional and Theoretical Economics*, *160*, 377–401.

Frey, B. S., and Osterloh, M. (2002). *Successful management by motivation: Balancing intrinsic and extrinsic incentives*. Berlin: Springer.

Frey, B. S., and Stutzer, A. (2002). *Happiness and economics: How the economy and institutions affect human well-being*. Princeton: Princeton University Press.

Gächter, S., and Fehr, E. (1999). Collective action as a social exchange. *Journal of Economic Behavior and Organization*, *39*, 341–369.

Gold, M. (1999). *The complete social scientist: A Kurt Lewin reader*. Washington, DC: APA.

Goodin, R. E. (1996). *The theory of institutional design*. Cambridge: Cambridge University Press.

Green, D. P., and Shapiro, I. (1994). *Pathologies of rational choice theory*. New Haven: Yale.

Harrison, L. E., and Huntington, S. P. (2000). *Culture matters: How values shape human progress*. New York: Basic Books.

Hastie, R., and Dawes, R. M. (2001). *Rational choice in an uncertain world*. Thousand Oaks, CA: Sage.

Higgins, E. T., and Kruglanski, A. W. (2001). Motivational science: The nature and functions of wanting. In E. T. Higgins and A. W. Kruglanski (Eds.), *Motivational science: Social and personality perspectives* (pp. 1–20). New York: Psychology Press.

Hogarth, R. (1980). *Judgment and choice*. New York: Wiley.

Hogg, M. A., and Abrams, D. (1988). *Social identifications*. New York: Routledge.

Kahneman, D., Slovic, P., and Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

Kahneman, D., and Tversky, A. (2000). *Choices, values, and frames*. Cambridge: Cambridge University Press.

Katz, H. C., Kochan, T. A., and Weber, M. R. (1985). Assessing the effects of industrial relations systems and efforts to improve the quality of working life on organizational effectiveness. *Academy of Management Journal*, *28*, 519–531.

Kelman, H. C. (1958). Compliance, identification, and internalization: Three processes of attitude change. *Journal of Conflict Resolution*, *2*, 51–60.

———. (1961). Processes of opinion change. *Public Opinion Quarterly*, *25*, 57–78.

———. (2006). Interests, relationships, identities: Three central issues for individuals and groups in negotiating their social environment. *Annual Review of Psychology*, *57*, 1–26.

Kelman, H. C., and Hamilton, V. L. (1989). *Crimes of obedience*. New Haven: Yale.

Koys, D. J. (2001). The effects of employee satisfaction, organizational citizenship behavior, and turnover on organizational effectiveness: A unit-level, longitudinal study. *Personnel Psychology*, *54*, 101–114.

Kramer, R. M. (1999). Trust and distrust in organizations. *Annual Review of Psychology*, *50*, 569–598.

Kramer, R. M., and Tyler, T. R. (1996). *Trust in organizations*. Thousand Oaks: Sage.

Lind, E. A., and Tyler, T. R. (1988). *The social psychology of procedural justice*. New York: Plenum.

McGuire, W. J. (1969). The nature of attitudes and attitude change. In G. Lindzey and E. Aronson (Eds.), *Handbook of social psychology* (2nd ed., Vol. 3, pp. 136–314). Reading, MA: Addison-Wesley.

Nisbett, R., and Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall.

Pfeffer, J. (1994). *Competitive advantage through people*. Cambridge, MA: Harvard.

Plous, S. (1993). *The psychology of judgment and decision making*. New York: McGraw-Hill.

Podsakoff, P. M., Ahearne, M., and MacKenzie, S. B. (1997). Organizational citizenship behavior and the quantity and quality of work group performance. *Journal of Applied Psychology*, *82*, 262–270.

Podsakoff, P. M., Bommer, W. H., Podsakoff, N. P., and MacKenzie, S. B. (2005). Relationships between

leader reward and punishment behavior and subordinate attitudes, perceptions, and behaviors. *Organizational Behavior and Human Decision Processes*, *98*, 113–142.

Shore, L. F., Barksdale, K., and Shore, T. H. (1995). Managerial perceptions of employee commitment to the organization. *Academy of Management Journal*, *38*, 1593–1615.

Skogan, W., and Frydl, K. (2004). *Fairness and effectiveness in policing: The evidence*. Washington, DC: National Research Council of the National Academy.

Stutzer, A., and Lalive, R. (2004). The role of social work norms in job searching and subjective well-being. *Journal of the European Economic Association*, *2*(4), 696–719. doi:10.1162/1542476041423331

Sunshine, J., and Tyler, T. R. (2003). The role of procedural justice and legitimacy in shaping public support for policing. *Law and Society Review*, *37*(3), 555–589.

Thaler, R. H. (1991). *Quasi-rational economics*. New York: Russell Sage Foundation.

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.

———. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453–458.

Tyler, T. R. (1994). Psychological models of the justice motive. *Journal of Personality and Social Psychology*, *67*, 850–863.

———. (2000). Social justice. *International journal of psychology*, *35*, 117–125.

———. (2006a). Legitimacy and legitimation. *Annual Review of Psychology*, *57*, 375–400.

———. (2006b). *Why people obey the law*. Princeton: Princeton University Press.

———. (2007). *Psychology and the design of legal institutions*. Nijmegen, the Netherlands: Wolf Legal Publishers.

———. (in press-a). Legitimacy and criminal justice: The benefits of self-regulation. *Ohio State Journal of Criminal Justice*.

———. (in press-b). *Why people cooperate*. Princeton: Princeton University Press.

Tyler, T. R., and Blader, S. L. (2000). *Cooperation in groups*. Philadelphia: Psychology Press.

———. (2005). Can businesses effectively regulate employee conduct?: The antecedents of rule following in work settings. *Academy of Management Journal*, *48*, 1143–1158.

Tyler, T. R., Boeckmann, R. J., Smith, H. J., and Huo, Y. J. (1996). *Social justice in a diverse society*. Boulder: Westview.

Tyler, T. R., Callahan, P., and Frost, J. (2007). Armed, and dangerous(?): Can self-regulatory approaches shape rule adherence among agents of social control. *Law and Society Review*, *41*, 457–492.

Tyler, T. R., and Darley, J. (2000). Building a law-abiding society: Taking public views about morality and the legitimacy of legal authorities into account when formulating substantive law. *Hofstra Law Review*, *28*, 707–739.

Tyler, T. R., Degoey, P., and Smith, H. (1996). Understanding why the justice of group procedures matters: A test of the psychological dynamics of the group-value model. *Journal of Personality and Social Psychology*, *70*, 913–930.

Tyler, T. R. and Fagan, J. (2008). Legitimacy and cooperation: Why do people help the police fight crime in their communities? *Ohio State Journal of Criminal Law*, *6*, 231–275.

Tyler, T. R., and Huo, Y. J. (2002). *Trust in the law*. New York: Russell-Sage.

Tyler, T. R., and Lind, E. A. (1992). A relational model of authority in groups. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 25, pp. 115–191). San Diego, CA: Academic Press.

Tyler, T. R. and Rankin, L. (in press-a). The mystique of instrumentalism. In J. Hanson (Ed.), *Ideology, psychology and law*. Oxford: Oxford University Press.

———. (in press-b). Public attitudes and punitive policies. In J. Dvoskin, J. L. Skeem, R. W. Novaco, and K. S. Douglas (Eds.), *Applying social science to reduce violent offending*. Oxford: Oxford University Press.

Tyler, T. R., and Smith, H. J. (1999). Sources of the social self. In T. R. Tyler, R. Kramer, and O. John (Eds.), *The psychology of the social self* (pp. 223–264). Hillsdale, N.J.: Erlbaum.

Walster, E., Walster, G. W., and Berscheid, E. (1978). *Equity*. Boston: Allyn and Bacon.

CHAPTER 5

Rethinking Why People Vote

Voting as Dynamic Social Expression

TODD ROGERS

CRAIG R. FOX

ALAN S. GERBER

In political science and economics, voting is traditionally conceived as a quasi-rational decision made by self-interested individuals. In these models citizens are seen as weighing the anticipated trouble they must go through in order to cast their votes, against the likelihood that their vote will improve the outcome of an election times the magnitude of that improvement. Of course, these models are problematic because the likelihood of casting the deciding vote is often hopelessly small. In a typical state or national election, a person faces a higher probability of being struck by a car on the way to his or her polling location than of casting the deciding vote. Clearly, traditional models cannot fully explain why and under what conditions citizens tend to vote.

In this chapter we will develop a novel framework for understanding why people vote. Instead of conceptualizing voting as a self-interested decision that is made at a single moment in time, we conceptualize voting as self-expressive social behavior that is influenced by events occurring before and after the actual moment of casting a vote. This conceptualization has several benefits. First, it helps to explain existing behavioral research that does not parsimoniously fit within the more traditional models of voting. Second, it helps identify several additional, currently underappreciated, factors that may affect people's likelihoods of voting. These derive from behavioral research in fields that have not previously been linked to voting (notably, social and cognitive psychology and behavioral economics).

Our conceptualization is best appreciated when viewed in contrast to traditional accounts of voting behavior. As described above, those conceive of voters as quasi-rational agents who evaluate whether to cast a vote by weighing the expected subjective benefit of voting against the expected subjective cost of voting. Those accounts generally encompass two types of benefits. The first is the impact that one expects her vote to have on the outcome of a given election. This "instrumental" benefit equals the difference in utility that a voter would derive from the preferred candidate versus the alternative candidate winning the election, multiplied by the subjectively assessed likelihood of casting the pivotal vote (Downs, 1957; Tullock, 1968). However, instrumental benefit cannot explain why millions vote in elections that they can reasonably be expected to know are not close. This fact gives rise to a "consumption" benefit from voting (Blais, 2000), which includes the pleasure a person experiences from fulfilling her civic duty to vote (Riker and Ordeshook, 1968) and the avoidance of the potential displeasure of having failed to vote when it might have mattered (Fiorina, 1974). The sources of this consumption benefit from voting have not been systematically analyzed. In some respects, the following account of voting as *dynamic social expression*—could be interpreted as unpacking this consumption benefit. However, not all of what we will describe fits neatly into this classification, and not all potential components of consumption benefit will be incorporated into our account.

We begin with a review of recent field experimental research exploring the impact of different modes of get-out-the-vote (GOTV) contact on turnout. The broad conclusion of this research is that the more personal the mode of contact, the more effective it is. Traditional models of why people vote are mostly silent on whether and why this would be the case. This deficiency helps motivate our conceptualization of voting as a fundamentally social behavior. In addition, we add two behavioral observations to our framework: voting is influenced by actions occurring before and after the moment of voting, and voting is an expression of one's identity.

We cite GOTV research throughout this chapter as empirical support for our account of why people vote. That research enables us to develop a behavioral

model grounded in observations of actual behavior, rather than a purely theoretical model grounded in questionable assumptions about behavior. In some cases, extant GOTV research confirms that parts of our model actually do causally affect people's likelihood of voting. In other cases, for which no extant research exists, we propose new GOTV field experiments to test our hypotheses. GOTV research is important not only for theoretical reasons but also for practical reasons: it can generate useful prescriptive insights for (cost-effectively) stimulating turnout in elections. The economic benefits of increased efficiency in GOTV efforts are significant because tens of millions of dollars are spent on such efforts in each federal election cycle. More importantly, increased effectiveness of GOTV efforts can achieve the social objective of increasing the number of citizens who participate in elections.[1]

The chapter is organized as follows. The first section, "Mode of GOTV Contact," reviews recent field experimental research exploring the impact of different modes of GOTV contact. This research helped motivate our conceptualization of voting as a fundamentally social behavior. Each of the three key elements of our framework of why people vote are then discussed. The second section, "Dynamic: Voting Is Affected by Events before and after the Decision," describes research supporting the view that voting behavior can be affected by actions occurring before and after the moment of actually casting a vote. The third section, "Social: Voting Is Influenced by Affiliative and Belonging Needs," discusses the implications of construing voting as a fundamentally social behavior. The fourth section, "Identity: Voting as an Expression of Identity," explores the potential implications of thinking of voting as an expression of one's personal and social identity. In the second through fourth sections, we discuss promising directions for future research to test and extend our conceptualization of why people vote. Finally, in the summary, we conclude with a brief review and discussion of our hopes for future research and theory building on this rich topic. Throughout this chapter we deliberately constrain our discussion to methods of promoting, rather than suppressing, participation, and to GOTV tactics that can be employed without the use of deceptive messages.

Mode of GOTV Contact: More Personal = More Effective

The last decade has witnessed an explosion in experimental field research examining the factors that influence citizens' likelihoods of voting. This began with the seminal 1998 study examining the relative impact of different modes of GOTV contact in an election in New Haven, Connecticut (Gerber and Green, 2000a). These investigators varied both the mode of contact and the content delivered to the citizens once they were contacted. Gerber and Green found no statistically meaningful differences in turnout across the variations in message content that they tested, but they did find very large differences across modes of communication. The vast majority of subsequent research in this area has continued to focus on the impact of different modes of GOTV contact, rather than on GOTV content. Thus, until recently little progress has been made in determining which GOTV message strategies are most effective in turning out voters. In fact, in Green and Gerber's quadrennial literature review summarizing the experimental work on voter mobilization, they underscore that the GOTV content does not seem to matter much (2004, p. 36).[2]

In the days and weeks leading up to an election, campaigns and their agents use a variety of communication modes to encourage citizens to vote. These range from the highly personal, such as face-to-face canvassing, to the highly impersonal, such as a prerecorded message over the phone ("robo-calling"). As mentioned, research in this area has generally found that the more personal the mode of contact, the larger its impact on the contacted citizen (Green and Gerber, 2004, 2008). In fact, Gerber and Green (2000a) suggest that the decline in voter turnout in the latter half of the twentieth century might be explained to a large extent by the use of increasingly impersonal modes of GOTV contact.

Personal, Face-to-Face Contact

Naturally, different forms of GOTV communication vary in their cost per contacted household: reaching a person through paid face-to-face canvassing is generally more expensive and labor-intensive than reaching a person through paid phone banking, and reaching a person through paid phone banking is generally more expensive and labor-intensive than reaching a person through paid direct mail.[3] That said, the mode of GOTV contact that results in the largest increase in turnout per contacted voter is personal, face-to-face contact. Initial experimental research found that a nonpartisan face-to-face canvassing effort had a 5–8 percentage point mobilizing effect in an uncontested midterm election in 1998 (Gerber and Green, 2000a) compared to less than a 1 percentage point mobilizing effect for live phone calls and for mailings. More than three dozen subsequent experiments have overwhelmingly supported the original finding that personal, face-to-face contact is more effective than less personal channels. The relative effectiveness of canvassing has been replicated in municipal elections (Arceneaux, 2005; Green, Gerber, and Nickerson 2003;

Michelson, 2003, 2005; Nickerson 2005) and federal elections (Arceneaux and Nickerson, 2006; Middleton and Green, 2008; Murray and Matland, 2005; Nickerson, Friedrichs, and King, 2006). It has also been demonstrated in several targeted populations, including younger citizens (Nickerson, 2006c; Nickerson, Friedrichs, and King, 2006), Latinos (Michelson, 2003; Murray and Matland, 2005), and African-Americans (Arceneaux, 2005).

Studies have looked at the effectiveness of canvassing efforts during low-salience elections (Arceneaux, 2005) as well as higher-salience, competitive elections (Bennion, 2005). Middleton and Green (2008) examined the canvassing effort of the partisan organization MoveOn.org in an especially high-salience election: battleground states during the 2004 presidential election. Uniquely, MoveOn.org's effort relied on local volunteers who were embedded in neighborhood social networks to mobilize voters. Face-to-face canvassers who are local and familiar can deliver GOTV contacts in especially personal ways compared to typical volunteers who are strangers to their GOTV targets. Impressively, this personalized form of canvassing resulted in a 9 percentage point increase in turnout compared to precincts that were not canvassed and were matched after the election based on identical observable characteristics. This impact is especially large when considering the very high salience of that election, and therefore the high level of baseline turnout.

Personal, Phone Calls

Dozens of experiments have examined the effectiveness of GOTV messages delivered by telephone. Several general findings emerge, all of which are consistent with the broad conclusion that the more personal a GOTV strategy, the more effective. First, the most effective phone calls are conducted in an unhurried, "chatty" manner. This has been found using professional phone banks especially trained to conduct conversational, unhurried calls (Nickerson, 2007) and using volunteers with training and good supervision (Nickerson, 2006d). Second, although even calls delivered in a rushed manner tend to have some effect (estimates vary but these calls appear to boost turnout by about 1 percentage point), they tend to be less effective than unhurried calls. This has been found using professional phone banks (McNulty, 2005; Nickerson, 2007) and volunteers (Nickerson, Friedrichs, and King, 2006). Finally, there is some preliminary evidence that recontacting those who had reported after an initial call that they intended to vote can be among the most effective phone-based GOTV methods (Michelson, McConnell, and Bedolla, 2009). We consider this strategy among the most "personal" of phone techniques because it involves referencing details of a past call. As we will discuss in "Dynamic Voting," this strategy also leverages the behavioral tool of *self-prediction and commitment.*

Impersonal, One-Way Communications

The least personal and the least effective GOTV communication channels entail one-way communications. First, written pieces encouraging people to vote that are mailed directly to households have consistently been shown to produce a small, but positive, increase in turnout (Green and Gerber, 2008). However, as we will see in the section "Identity," a recent study has suggested that more personalized content included within the direct mail pieces (e.g., showing citizens their voting record and that of their neighbors) can render them much more effective (Gerber, Green, and Larimer, 2008). Second, GOTV leaflets delivered to households by canvassers have been found to have small positive effects on participation rates (Gerber and Green, 2000b; Nickerson, Friedrichs, and King, 2006), especially among unaffiliated voters (Gerber and Green, 2000b). Third, calling households to deliver a prerecorded script, what are known as "robocalls," has not been found to have any measurable impact on turnout (Green and Gerber, 2008). Finally, GOTV email messages have no effect whether sent by partisan organizations (Stollwerk, 2006) or nonpartisan organizations (Nickerson, 2006a). All told, these impersonal modes of contact have a small-to-negligible effect in stimulating participation.

Interpreting the Impact of More Personal Communications

Why are more personal modes of GOTV contact more effective in stimulating turnout? Traditional rational models of voter behavior might suggest the following answers. First, personal modes of GOTV contact may have more impact because they affect how likely a citizen would be to notice the information (e.g., it is easier to dismiss a message presented on a leaflet than a message delivered by a person at one's door). Second, citizens may more carefully attend to messages delivered in more personal and interactive ways (e.g., a person may listen more intently to and engage with a message delivered by a person at their door than with a message delivered by mail).

Though enhanced attention no doubt contributes to the heightened impact of more personal communications, we suggest that this heightened impact is enhanced by the social dimension of more personal interactions. For instance, the attention account cannot readily explain why even the most effective telephone calls are less than half as effective as face-to-face canvassing. Apparently, some aspect of face-to-face

interactions renders targets more receptive to appeals (Reams and Ray, 1993). Naturally, a deeper social connection is fostered in face-to-face interaction than over a telephone. This social connection likely engages people's empathy and their fundamental desire for acceptance, both of which tend to engage the motivation to behave in socially desirable ways (Baumeister and Leary, 1995). Additionally, more personal communication channels facilitate the detection of social similarity between the target and the communicator, which has been shown to increase a target's likelihood of complying with requests (Burger et al., 2004). Finally, more personal GOTV communication may provide an opportunity for targets to make more compelling commitments about their future behavior. Indeed, asking people to publicly commit to a future behavior (e.g., voting) has been shown to increase their likelihood of following through on that behavior (e.g., Sherman, 1980), and such commitments have greater impact when they are made in more public ways (Deutsch and Gerard, 1955).

Mode of GOTV Contact: Summary and Future Directions

Gerber and Green's initial New Haven experiments, and the many experiments that have followed, developed a method for assessing the effectiveness of GOTV communication channels and inductively accumulated insights into what motivates people to vote. Future research should explore other modes of GOTV contact such as television or radio ads encouraging turnout (surprisingly little of which has been done to date) and the common practice of holding signs on highly trafficked streets to remind people of an election. Other modes of GOTV contact include emerging digital technologies, such as online banner ads, social networking tools like Facebook or Twitter, and text messaging. Preliminary research on some of these modes of contact has already begun (TV: Gerber et al., 2006; Green and Vavreck, 2006; radio: Panagopoulos and Green, 2006; Internet: Iyengar, 2002; text messaging: Dale and Strauss, 2007), and a clearer understanding of their effectiveness will be of substantial value in the years to come. Another important factor affecting a citizen's likelihood of voting is his or her eligibility to cast a vote. Eligibility involves individual-level registration status (naturally, today's unregistered voters are less likely to vote in this fall's election), state-level registration rules (for instance, how cumbersome a process is required), and state-level voting qualification requirements. All three of these are rich areas for additional research that could inform GOTV best practices, laws regarding election eligibility, and, most fundamentally, our understanding of why people vote. Finally, it merits mention that voting early, either by mail or in person, is increasingly popular. A better understanding of how to specifically mobilize citizens to vote early, and the impact of early voting on overall turnout, will be extremely valuable. For example, in the 2008 U.S. presidential election, 24.3% of total votes cast were cast before election day. While initial research suggests that encouraging early voting might increase turnout (Mann, 2009), many questions remain unanswered.

As mentioned, the traditional account conceives of voting as a static, self-interested, and quasi-rational decision. Such models cannot readily accommodate the experimental findings that more personal modes of GOTV contact are more effective in mobilizing citizens to vote. To accommodate the impact of communication mode on voter mobilization as well as new findings concerning the impact of specific messages, we propose to modify the traditional account of voting in three respects. First, we note that voting is not merely a static event that occurs at a single point in time but rather a dynamic constellation of behaviors that are extended over time, from the initial formation of an intention to vote to the act of casting a vote to the subsequent knowledge that one has or has not voted. Second, voting is not a purely self-interested act but an inherently social one that may accrue not only instrumental and consumption benefits but also fulfill basic needs of affiliation and belonging to a larger group. Third, voting is not merely a decision, it is also an expression of one's identity. Conceiving of voting as a dynamic social expression broadens the range of factors that can influence voting in three important respects. The following three sections will explore some implications of each of these facets in turn.

Dynamic: Voting Is Affected by Events before and after the Decision

Conceiving of voting not as a static decision but rather as a constellation of behaviors that extend over time suggests that events that occur before and after the moment when a person decides to vote can affect whether or not she actually follows through and casts a vote. In this section we will discuss two areas of behavioral research that are relevant to what occurs before the moment a person decides whether or not to vote. We will then discuss a third area of behavioral research that is relevant to what occurs after the moment a person decides whether or not to vote.

Before Deciding to Vote: Self-Prediction and Commitment

One means to facilitate the performance of a socially desirable behavior is to ask people to predict whether

they will perform the behavior in the future. In order to present oneself in a favorable light or because of wishful thinking or both, people are generally biased to answer in the affirmative. Moreover, a number of studies have found that people are more likely to follow through on a behavior after they have predicted that they will do so, a pattern referred to in different literatures as the "self-erasing nature of errors in prediction" (Sherman, 1980), the "self-prophecy effect" (Greenwald et al., 1987), and the "mere measurement effect" (Morwitz, Johnson, and Schmittlein, 1993). In one classic study (Sherman, 1980) participants were contacted over the phone to answer questions about a variety of topics. For half of participants, the survey included the question of whether they believed they would volunteer for three hours at the American Cancer Society if they were ever asked to do so; 48% of these participants said they thought they would. The other half of participants were not asked to make such a prediction. Three days later a different volunteer came to all participants' doors to ask if they would volunteer for the American Cancer Society. Whereas only 4% of participants who had not made a self-prediction agreed to volunteer, a whopping 31% of participants who had previously made a self-prediction agreed to volunteer. Thus, participants were optimistic in predicting their likelihood of agreeing to volunteer, but the act of making a public affirmative prediction made them substantially more likely to volunteer than had they not self-predicted.

Several factors have been found to moderate the effect of self-prediction on behavior. First, the effect is stronger when people turn their self-predictions into commitments, articulating a desire and a will to perform the behavior. Commitment elicitation adds a socially binding element to self-prediction and increases the social costs of failing to fulfill one's self-prediction (Schlenker, Dlugolecki, and Doherty, 1994). Self-commitment has been found to increase compliance even in the absence of explicit accountability (for a review see Cialdini, 2003). This is because commitments activate a basic desire in people to bring behaviors into consistency with their beliefs and their expectations about themselves (Bem, 1972; Festinger, 1964). Second, self-prediction/commitment effects tend to be much stronger when they are made in more public ways. For instance, one study found that a public vote makes three-person juries more likely to deadlock (Kerr and MacCoun, 1985). Third, self-predictions/commitment effects are stronger when they are viewed as authentic and voluntary, and they tend to diminish or disappear to the extent that they appear to be the result of bribery or coercion. For instance, in one classic study, participants asked to tell another student that a boring task had been fun were more likely to rate the task as actually having been interesting if they had been paid a paltry $1 to talk up the study than if they had been paid a coercive $20 to do so (Festinger and Carlsmith, 1959).

Self-prediction/commitment effects have an obvious application to the GOTV context: asking citizens to articulate their intention to vote should make them more likely to actually turnout. In fact, one of the earliest studies examining the effect of self-prediction on behavior examined the domain of turnout: asking a small number of college students if they intended to vote in the 1984 U.S. general election and asking them to verbalize their most important reason for voting increased their likelihood of actually voting by more than 23 percentage points (Greenwald et al., 1987). A confound in the design was that the treatment did not isolate self-prediction, but instead also included a question about why voting is important. Given that these two questions were combined for the study's treatment group, one cannot be certain whether self-prediction, specifically, caused the increased turnout. Greenwald and colleagues ran a follow-up study in which they attempted to replicate the original finding and to isolate the effect of self-prediction (Greenwald et al., 1988). The follow-up experiment found a comparably sized self-prediction effect, but only among participants whose vote history suggested that they were occasional voters, as opposed to those who had consistently voted or failed to vote. They also found that the follow-up question regarding why people think they should vote had an additive effect, but also only among occasional voters.

Some caveats are in order when translating these studies into a contemporary GOTV context. The original studies were conducted over the phone more than two decades ago, when telephone calls were less widely used for voter mobilization. Recent election cycles have seen an increased use of the telephone as a GOTV communication channel, and we suspect that this practice could result in the decreased efficacy of any single call relative to those made in the 1980s. In a more recent study conducted during the 2000 U.S. presidential primary (Smith, Gerber, and Orlich, 2003), 1,160 citizens were contacted and assigned to one of four conditions: control, self-prediction only, reason-to-vote only, combined self-prediction and reason-to-vote. All conditions followed the procedures used by Greenwald and his collaborators. In contrast to the very large treatment effects reported by them, this experiment found very small, statistically insignificant treatment effects for self-prediction or for eliciting a reason to vote. Smith and colleagues also failed to find the effect for occasional voters relative to frequent and infrequent voters suggested by the follow-up study by Greenwald and colleagues (1988).

Smith, Gerber, and Orlich (2003) suggested that the effect sizes found by Greenwald and his colleagues

(1987, 1988) may not replicate in a contemporary GOTV application. However, the Smith study does not altogether disprove that self-prediction can be a useful part of GOTV content—exploratory analysis merging the infrequent and occasional voters together reveals that the self-prediction treatment (including all participants who made a self-prediction) resulted in a (nonsignificant) 3.2 percentage point increase in turnout. As we will discuss later, these two subgroups appear to be the most susceptible to other behavioral interventions as well (see "Following the Herd: Descriptive Social Norms," below). A recent study conducted during the 2008 presidential primary in Pennsylvania (N = 287,228) found that GOTV election encouragement calls that also elicited a self-prediction resulted in a marginally significant 2.0 percentage point increase in turnout relative to an untreated control group (Nickerson and Rogers, 2010).

Future research can examine different modes of eliciting self-prediction and commitment and can also contribute to our knowledge of the underlying mechanisms. For example, pledges or petitions could be incorporated into canvassing efforts or rallies. Emerging social-networking technologies provide new opportunities for citizens to commit to each other that they will turnout in a given election. These tools facilitate making one's commitments public, and they also allow for subsequent accountability following an election (see "Thinking about What Happens after the Election: Social Pressure and Accountability," below). In addition to demonstrating the relevance of this behavioral phenomenon in the domain of voting, research on this topic could also advance the basic behavioral science. For example, it could address questions such as whether commitment and self-prediction become more or less effective when leveraged repeatedly (e.g., over several election cycles). Also it is an open question whether citizens become more accurate, and less optimistically biased, in their self-predictions when asked multiple times over several elections.

After Deciding to Vote: Implementation Intentions

Though public self-predictions and commitments have been found to increase the likelihood that people follow through on what they say they will do, behavioral research has identified an even more effective method for increasing that tendency. Asking people to form a specific if-then plan of action, or implementation intention, reduces the cognitive costs of having to remember to pursue an action that one intends to perform (Gollwitzer, 1999; Gollwitzer and Sheeran, 2006). Research shows that when people articulate the how, when, and where of their plan to implement an intended behavior, they are more likely follow through. This occurs by cognitively linking two critical elements. First, by anticipating a situation that is important for implementing an intention (e.g., locating one's polling place), one is more likely to automatically recognize in the moment that the situation is critical for fulfilling one's intention (e.g., register one's vote). Second, by anticipating how one will behave in a given situation (e.g., on my way to work next Tuesday morning), one is more likely to automatically behave in ways that fulfill one's intention. Implementation intentions link intention-relevant situations with intention-relevant behaviors. These pairs can be thought of as "if situation Y, then behavior X" (Gollwitzer, Bayer, and McCulloch, 2005; Gollwitzer and Sheeran, 2006).

The formation of implementation intentions has been shown to affect dozens of repeated behaviors, such as taking vitamin supplements (Sheeran and Orbell, 1999) or exercising (Lippke and Ziegelmann, 2002; Milne, Orbell, and Sheeran, 2002). An aspect that is more relevant to voting is that implementation intentions have also been found to increase the likelihood of completing one-time behaviors that must be executed within a finite window. In one study, students were told to pick up their reading material at their teaching assistant's office during an eight-hour window on the following day (Dholakia and Bagozzi, 2003, study 1). Half the participants were told that the materials were optional for the course, but they were instructed to formulate detailed implementation intentions about when, where, how, and how long it would take to pick up the reading materials at the TA's office. The other half of participants were told that the readings were very important for the course, but they were not instructed to formulate implementation intentions. Results showed a dramatic effect of the manipulation: a large majority of students in the implementation intentions condition (72%) actually retrieved the reading materials during the eight-hour window the following day, whereas a minority of the students who were merely told that the materials were very important (43%) actually retrieved them during the specified window.

Translating research on implementation intentions into the GOTV context would first entail eliciting from citizens a goal intention to vote. Notice that goal intentions are self-predictions and thus exploit the aforementioned self-prediction effect, if one occurs. Second, translating implementation intentions into the GOTV context would involve prompting citizens to detail how they will follow-through on their goal intention to vote. When will they vote? How will they get to their polling place? Where will they be before going to their polling place? One as-

pect of facilitating implementation intentions that is especially appealing for GOTV efforts is that it could be incorporated into the GOTV telephone calls preceding an election that are currently in widespread use by campaigns. A recent experiment conducted during the 2008 presidential primary in Pennsylvania (N = 287,228) found that GOTV phone calls eliciting implementation intentions increased turnout by 4.1 percentage points relative to an untreated control group. This treatment effect was more than twice as great as an election encouragement call that also elicited a self-prediction (Nickerson and Rogers, 2010). More research is needed, but these are promising initial findings.

Thinking about What Happens after the Election: Social Pressure and Accountability

Conceptualizing voting as dynamic behavior rather than a static decision suggests that events that occur after the decision to vote, and even after the act of voting, can affect one's likelihood of voting. In particular, holding a person publicly accountable for whether or not she voted may increase her tendency to do so. Feelings of accountability can be induced by leading people to believe that they could be called upon to justify their behavior to others after making a judgment, decision, or performing an action (see Lerner and Tetlock, 1999). Studies have found that when people are merely made aware that their behavior will be publicly known, they become more likely to behave in ways that are consistent with how they believe others think they should behave (Posner and Rasmusen, 1999; Rind and Benjamin, 1994). Accountability has been successfully leveraged in public campaigns to pressure people to perform socially valued behaviors. For instance, at one point Italy exposed those who failed to vote by posting the names of nonvoters outside of local town halls (Lijphart, 1997: 9 n18).

In a recent field experiment, Gerber, Green, and Larimer (2008) investigated the effectiveness of manipulating accountability in a direct-mail message. A first group of households received a mailing with a message encouraging them to vote. A second group of households received a similar mailing with the additional information that researchers would be studying whether or not the residents of the household voted by examining publicly available records. This condition tested the effect of having one's voting behavior observed by a third party, in this case anonymous researchers. A third group of households received a similar mailing in which the message also included a display of the turnout history of those who reside in the household. This message also reported that a follow-up letter would be sent after the upcoming election to report who in the household voted and who did not. This condition tested the effect of having one's voting behavior known to others in one's household. Finally, a fourth group of households received a similar mailing in which the message included a display of the turnout history of not only those who reside in the household, but also that of their neighbors. This mailing also reported that one's neighbors received a similar mailing, and that the recipient and his or her neighbors would receive a follow-up letter after the election to show who in the neighborhood had voted and who had failed to vote. This condition tested the effect of having one's voting behavior known to others in one's neighborhood, in addition to one's own household. Altogether, this study examined the effect of varying degrees of accountability induced by a single mail piece on citizen's voting behavior.

Results indicated a dramatic impact of the social accountability manipulation on turnout: the condition that induced the greatest level of social accountability (in which one's neighborhood was involved) resulted in an astonishing 6.3 percentage point increase in turnout compared to the mailing that used the standard encouragement to vote message. This study demonstrates that a normally impersonal and ineffective GOTV channel (direct mail) can be used to deliver a highly personalized message that strongly impacts turnout. To put this in context, a standard GOTV mailing has around a 0–2 percentage point impact on turnout (Green and Gerber, 2008).

Social: Voting Is Influenced by Affiliative and Belonging Needs

The second facet of our conceptual model of voting as dynamic social expression is that it is a fundamentally social act. People are strongly motivated to maintain feelings of belonging with others and to affiliate with others (Baumeister and Leary, 1995). Failure to meet these needs can have consequences for health (Kiecolt-Glaser et al., 1984; Lynch, 1979), and well-being (Lyubomirsky, Sheldon, and Schkade, 2005; Myers, 1992). The insight that voting can partly satisfy these social needs can generate a number of GOTV content strategies. We have already mentioned the effectiveness of manipulating social accountability concerning whether or not a person casts a vote. Other GOTV strategies that can increase turnout by serving social needs could involve encouraging people to go to their polling place in groups (i.e., a buddy system), hosting after-voting parties on election day, or encouraging people to talk about voting with their friends, to name a few. In this section we will describe behavioral research that explores some GOTV strategies motivated

by the insight that people are concerned for others and that they tend to behave in ways that are consistent with social expectations of appropriate behavior.

Tending to One's Own: Voting for the Sake of Others

Social identity theory posits that people spontaneously classify themselves and others into groups. People derive self-esteem from their membership with groups, even if those groups are arbitrary or ad hoc (Tajfel, 1982). Once people identify with an in-group, they are willing to incur a cost to help other members of their group. For instance, one study looked at people's willingness to give up money for the sake of a member of one's group in a dictator game, a strategic interaction in which one player (the "dictator") is asked to allocate a fixed amount of money between herself and another participant (Forsythe et al., 1994). Typically, studies of the dictator game have found that the average allocation of money from the dictator to anonymous others ranges from 10% to 52%, despite the fact that the rational solution is for the dictator to keep all of the money for himself (Camerer, 2003). Interestingly, Fowler and Kam, 2007 (see also Edlin, Gelman, and Kaplan, 2007; Fowler and Kam, 2006; Jankowski, 2002) found that people allocated more money to an anonymous participant who shared their political party identification than to an anonymous participant who had a different party identification.

Incorporating the welfare of others into why people vote can have several implications for stimulating turnout. In particular, messages that emphasize the importance of the issues at stake in the election for *other* favored citizens (e.g., one's neighbors, friends, or family) may motivate citizens to vote. Although this approach may seem obvious, it has not been used systematically in GOTV messaging and has not yet been studied carefully in controlled field experiments.

Following the Herd: Descriptive Social Norms

The basic need for belonging can influence people to behave in ways that are consistent with how they expect others to behave. This expectation is referred to as a descriptive social norm. Research by Cialdini and colleagues has found that people tend to conform to descriptive social norms, particularly when people feel uncertain about what kind of behavior is appropriate (Cialdini, 2003; Cialdini, Reno, and Kallgren, 1990; Reno, Cialdini, and Kallgren 1993). Note that the motivation to conform may be strong even if the descriptive social norm violates how others believe a person *should* behave (i.e., the prevailing injunctive norm). This research suggests that including descriptive social norms in persuasive appeals when actual behavior runs counter to a community's desired behavior can have perverse effects (Cialdini et al., 2006). If a descriptive social norm does not reflect a desired behavior (e.g., "The park is actually full of litter . . ."), then highlighting the descriptive social norm, even if to contrast it with the desired behavior (e.g., ". . . so please do not litter"), can actually impair the effectiveness of the appeal (Cialdini et al., 2006; Cialdini, Reno, and Kallgren 1990, experiment 1;). This is because in addition to saying "one should not litter," the message also says "many people do litter." Studies have demonstrated the strong influence of descriptive social norms on behavior in a variety of real-world situations, including littering (Cialdini, Reno, and Kallgren, 1990), recycling (Cialdini, 2003), binge drinking on college campuses (Mattern and Neighbors, 2004), stealing (Cialdini et al., 2006), and towel reuse in hotels (Goldstein, Cialdini, and Griskevicius, 2008).

Political campaigns often use descriptive social norms in GOTV content, but they sometimes do so in detrimental ways. For example, in the final days before the 2004 U.S. presidential election, when candidate John Kerry spoke to a group of women in Milwaukee, he referred to the "roughly 38 million women who didn't vote in 2000." We surmise that this approach is common among political professionals because they are not aware of the power of descriptive social norms. A survey of self-identified experts in GOTV confirms this suspicion: 43% reported believing that a message emphasizing that "turnout among the young is relatively low and/or decreasing" would be more effective in motivating turnout than another emphasizing that "turnout among the young is relatively high and/or increasing" (Rogers, 2005).

Although research from social psychology suggests that emphasizing high turnout will be more motivating than emphasizing low turnout, there are reasons why one might suspect this will not be the case in the context of voting. In particular, the higher the turnout is for a given election, the less likely any one person's vote will affect the outcome (Downs, 1957). Even if a voter were concerned with not only which candidates and issues prevail but also the margin of victory (e.g., to increase the mandate for the favored candidate or issue; see Light, 1999), a vote cast in a low-turnout election will be of greater political importance than a vote cast in a high-turnout election.

Recent research by Gerber and Rogers (2009) has examined the impact of descriptive social norms in two field experiments during statewide gubernatorial elections in New Jersey and California. Each experiment had the same general structure. Participants were called by a professional GOTV phone bank during the three days before the election and strongly en-

couraged to vote. Half of participants heard a message that used true statistics about turnout in elections over the previous twenty years to emphasize that turnout would be high in the upcoming election. These participants heard statements such as "In the last election [in CA or NJ] X million citizens VOTED." The remaining participants heard a message that used true statistics about turnout in elections over the previous twenty years to emphasize that turnout would be low in the upcoming election. These participants heard statements such as "In the last election [in CA or NJ] X million citizens FAILED TO vote." At the end of all messages, the strength of participants' motivation to vote in the upcoming election was elicited. Both studies showed that the motivation to vote significantly increased when participants heard a message that emphasized high expected turnout as opposed to low expected turnout. For example, in the New Jersey study, 77% of the participants who heard the high-turnout script reported being "absolutely certain" they would vote, compared to 71% of those who heard the low-turnout script. This research also found that moderate and infrequent voters were strongly affected by the turnout information, whereas frequent voters were unaffected by the information.

Identity: Voting as an Expression of Identity

The final facet of our account of voting as dynamic social expression is that citizens can derive value from voting through what the act displays about their identities. People are willing to go to great lengths, and pay great costs, to express that they are a particular kind of person. Consumer research has shown, for example, that brands that people purchase tend to be viewed as an extension of their identities (Belk, 1988; Fournier, 1998). Similarly, social identity research has shown that people are motivated to behave in ways that are consistent with behavior of in-groups with which they most strongly identify and that doing so boosts their self-esteem (Tajfel, 1982). They also strive to be seen by others as they see themselves (Swann and Read, 1981). Moreover, people experience dissonance when their behavior contradicts their beliefs; behaving in ways that are consistent with one's self-views can avoid this aversive state of dissonance (Festinger, 1964). For all of these reasons, the candidate or party for whom one votes and the very act of voting may serve important signaling functions to oneself and to others.

Conceiving of voting as an act of self-expression suggests at least three approaches to increasing voter turnout. First, one can influence how a citizen construes what it means to vote. Casting a vote could be framed as meaning anything from "I care about this election outcome" to "I care about my family's future and setting a good example for them" to "I care about my society and fulfilling my civic duty." GOTV content that emphasizes a meaning that is more highly valued by voters should be more effective at mobilizing voting.

A second way to increase the expressive value of voting is to increase the extent to which a citizen's voting behavior will be observed by other members of one's in-group. Recall the study mentioned in "Dynamic: Voting Is Affected by Events before and after the Decision" that found that voter turnout was increased by the threat to publicize a citizen's voting record after the upcoming election (Gerber, Green, and Larimer, 2008). In that section we highlighted the motivational power of the shame of nonvoting being exposed publicly. We also suspect that part of the motivational power of this intervention derives from the pride of having one's successful voting being publicly recognized. Such pride in voting can also be engendered in several other ways; for example, by providing to those who cast a vote stickers that say "I voted!" or by posting voting records in public places.

The third means of changing the expressive value citizens derive from voting is to influence the extent to which the act of voting expresses a desired identity. We will focus on this approach as we review research on three tactics shown to affect people's behavior by changing how they see themselves, and we will discuss how each might be employed in the GOTV context.

Initiating the "Voter Identity": Foot-in-the-Door

One common tactic that influences behavior by engaging a target's identity is the so-called foot-in-the-door technique. This tactic involves asking a person to accede to a relatively small request in order to increase the likelihood that he or she will agree to a larger request in a related domain in the future. For instance, in one classic study, this technique was used to increase the percentage of people willing to post a large, crudely written sign on their front lawns that read "DRIVE CAREFULLY" (Freedman and Fraser, 1966). Half of participants were asked by a stranger who came to their homes if they would be willing to display the sign. Only 17% agreed. The other half of participants had been approached by a different stranger two weeks earlier and asked if they would place a small, three-inch sign in their window or car that read "Be a safe driver." Nearly all agreed to this first minimal request. However, when these people were asked to post the large, crudely written billboard in their lawns, an astonishing 76% agreed. This surprising effect arose because participants who first

agreed to post the three-inch sign came to see themselves over the course of the two intervening weeks as "the kind of people who care about safe driving." The increased willingness to acquiesce to the subsequent bigger request (e.g., posting the large billboard in their lawns) has been interpreted as resulting from a change in the targets' perceptions of themselves.

In order for the foot-in-the-door technique to increase a behavior, several conditions must be met (Burger, 1999). First, people must interpret their original small behavior as having been of their own choosing and as not having been motivated by some other extrinsic reward (Festinger and Carlsmith, 1959). Second, the more often people are reminded of their original small behavior, the more effective it will be in influencing their self-perceptions (Hansen and Robinson, 1980). At the same time, however, there is a danger of making the first request so large that a person can decide that having performed it, she has "done enough" (Cann, Sherman, and Elkes, 1975; Snyder and Cunningham, 1975), especially if the first request was made by the same requester immediately before the second request (Chartrand, Pinckert, and Burger, 1999). Third, the first request must elicit a high percentage of acquiescence. Just as when people agree to a small behavior they become more likely to later agree to a larger behavior, if people do not agree to the first request, they may become *less* likely to later agree to a larger behavior.[4]

The foot-in-the-door method could be used in GOTV strategy by asking citizens to comply with a small request relevant to voting prior to election day. This could include wearing pins on their shirts, putting bumper stickers on their cars, or volunteering a small amount of time or money to a campaign. Thus far, use of the foot-in-the-door technique has not yet been well studied in the context of GOTV. However, one study attests to the power of such initial requests on subsequent voting: citizens who would not have voted in an odd-year local election but were induced to do so in a GOTV canvass experiment were almost 60% more likely to vote in the subsequent election the following year compared to citizens who were not induced to vote in the odd-year local election (Gerber, Green, and Shachar, 2003).

Voting as a Self-Fulfilling Prophecy: Identity Labeling

Identity labeling entails explicitly reinforcing a facet of a person's real or ideal self that is associated by the desired behavior. This could be a group identity (i.e., an American citizen) or a more personal self-categorization (i.e., the kind of person who cares about America) (Turner et al., 1987). For example, one study looked at the effect of creating and reinforcing in fifth-grade students the social identity that they were members of a litter-free classroom. The researchers reinforced this identity over the course of eight days. One example of how this was done is that on the fifth day a sign was posted in the room reading "We are [Mrs.] Andersen's Litter-Conscious Class." This social-identity reinforcement more than tripled the percentage of litter discarded in the wastebasket relative to that of a control classroom (Miller, Brickman, and Bolen, 1975). The treatment was also more than twice as effective as repeatedly asking a different set of students over a similar eight-day period not to litter.

The identity-labeling tactic could be factored into GOTV content in a variety of ways. One method would be to reinforce and make salient an identity that a person already likely possesses that would encourage her to vote. For example, one could develop a message that emphasizes a target's identity as an American, as a parent or grandparent, as an environmentalist, as a soldier, etc. This method would entail selectively reinforcing the preexisting identity that is most likely to induce the pro-social behavior of voting.

Another method would be to induce an identity that may not already exist but that is plausible. A common method used to do so in past research is to ask participants to complete a survey that is ostensibly intended to assess the degree to which the participants possess some characteristic. After completing the instrument the experimenter provides (false) feedback using a label that allegedly derives from participants' responses. A study in 1978 used this method to determine how potent identity labeling could be in voter mobilization (Tybout and Yalch, 1980). Experimenters asked participants to complete a fifteen-minute survey that related to an election that was to occur the following week. After completing the survey, the experimenter reviewed the results and reported to participants what their responses indicated. Participants were, in fact, randomly assigned to one of two conditions. Participants in the first condition were labeled as being "above-average citizen[s] . . . who [are] very likely to vote," whereas participants in the second condition were labeled as being "average citizen[s] . . . with an average likelihood of voting." Participants were also given an assessment sheet corresponding to their labels. These identity labels proved to have substantial impact on turnout, with 87% of "above average" participants voting versus 75% of "average" participants voting.

While this study provides insight into the potential of identity labeling for GOTV content, it must be interpreted with caution for several reasons. First, this study relied on a small sample size ($N = 162$) and has

not yet been replicated. Second, the study was conducted more than two decades ago, so the contemporary political environment may result in participants responding differently to such a design. Third, the fact that the average turnout across conditions was so high (81%) indicates that the population used in the study was prone to voting in the first place. This meant that the "above average" label was probably credible to those who received it. Such a label would likely not be credible when delivered to members of a population who rarely, or never, vote. In order for an identity label to be effective, it must be credible to its recipient (Allen, 1982; Tybout and Yalch, 1980).

The identity-labeling method used in this study is also ethically dubious because it depends upon delivering false or misleading feedback to participants. However, this technique could be ethically applied in a variety of ways, the simplest of which is to merely assert that a target citizen is the kind of person who values his or her right to vote. How best to use identity labeling to increase turnout is a promising avenue for future research. A first-order question is, Through what mode of GOTV contact can identity labeling have an impact? These include direct-mail pieces, television, radio, and billboard advertising, speeches, and all direct contact with potential voters (e.g., canvassing, rallies, etc.). While we surmise that the more personal and interactive modes of GOTV contact will enable the strongest identity-labeling treatments, it is not inconceivable that vivid mail or TV messages could be highly effective as well.

Seeing Oneself Voting: Visual Perspective

A third tactic for changing behavior by affecting how people see themselves involves using a visualization technique. Illustrations of this tactic build off classic research showing that actors and observers tend to have different explanations for behaviors. Whereas observers are prone to attributing a behavior they witness (i.e., a person tripping over a rock) to dispositional characteristics (i.e., the person is clumsy), actors tend to attribute the same behavior to situational factors (i.e., the trail was treacherous) (Gilbert and Malone, 1995; Jones and Nisbett, 1971). More recently, studies have found that when people are induced to recall their own past behavior from an observer's perspective, it increases their tendency to attribute their behavior to their own disposition, relative to when they are induced to recall their own past behavior from a first-person perspective (Libby, Eibach, and Gilovich, 2005).

In a recent study exploring how visual perspective can affect voting, Ohio college students were guided through a one-minute visualization that entailed picturing themselves entering the voting booth and casting a vote. This visualization took place on the night before the 2004 U.S. presidential election. One group of participants were guided to picture themselves from the third-person perspective, and another group of participants were guided to picture themselves from the first-person perspective. Three weeks after the election, the participants reported whether or not they had voted in the election: 90% of those who had been guided to visualize themselves voting from the third-person perspective reported having voted, whereas only 72% of those who had been guided to visualize themselves voting from the first-person perspective reported having voted. Moreover, the difference in reported turnout was statistically mediated by the extent to which participants reported seeing themselves as the kind of people who vote. Although this study had a small sample size ($N = 90$) and measured self-reported behavior rather than actual voting behavior, the tactic merits follow-up research. Like the previous two tactics for leveraging voting as an expression of identity, this one suggests a potentially powerful tool for stimulating turnout.

Summary and Conclusion

In this chapter we have observed that one challenge to traditional accounts of voting as a static, self-interested and quasi-rational decision is that voter mobilization efforts are more successful when communicated through more personal media. We have advanced an alternative account of voting as dynamic social expression. In motivating each facet of this reconceptualization we have drawn on behavioral research that has not been traditionally cited in the GOTV literature. Note that the three facets we discuss (dynamic, social, and expression of identity) are somewhat overlapping categories; for example, the social accountability intervention that we cited above relies on all three: (1) it works because people consider a consequence long after the decision to vote that has nothing to do with the election outcome (i.e., it is dynamic); (2) it works because people care how their neighbors view them (i.e., it is social); and (3) it works because people wish to see themselves as good citizens (i.e., it entails an expression of identity).

Of course, traditional models could be extended to accommodate these factors. For example, the positive influence on voting of articulating implementation intentions could be modeled as a reduction in the cognitive costs of voting. Similarly, satisfying affiliation needs by casting a vote could be modeled as a consumption benefit of voting. However, we assert that the power of our new conceptual model is that

it is theoretically generative: it makes explicit a set of new variables that have been found to empirically influence behavior (and often voting itself) that do not naturally follow from the traditional model of voting as a static self-interested decision.

We also wish to underscore the fact that not all citizens will respond equally to each of the behavioral interventions mentioned in this chapter; naturally, some people are more susceptible to some types of influence than others. In recent years GOTV professionals have found it effective to tailor "microtargeted" messages that highlight particular issues to specific individuals based on their consumption habits and demographic characteristics (Fournier, Sosnik, and Dowd, 2006; Gertner, 2004). Likewise, we suspect that the effectiveness of particular kinds of behavioral appeals might be predicted from observable demographic variables. For example, as was discussed above, Gerber and Rogers (2009) found that though infrequent and occasional voters were highly affected by whether or not expected turnout would be high or low, frequent voters were unaffected in either direction. This result suggests that GOTV content involving descriptive social norms should be targeted at voters who are expected to be moderately likely or unlikely to vote. Similarly, the study looking at the effect of self-prediction on subsequent voting (Smith, Gerber, and Orlich, 2003) suggested that this same subgroup of citizens might be most susceptible to self-prediction and commitment effects.[5]

Another example of a psychographic characteristic that could prove promising for microtargeting GOTV content is a person's propensity to self-monitor (Gangestad and Snyder, 2000; Snyder, 1974). Highly self-monitoring persons are especially concerned with how others see them. This characteristic has been shown to be positively related to how much people change their behavior when they are made aware that others will know how they behave in a given situation (Lerner and Tetlock, 1999; Snyder, 1974). High self-monitors tend to conform to what they believe they "should" do when they are aware that others will know about their behavior. One could imagine that the accountability intervention reported by Gerber, Green, and Larimer (2008) could be especially effective on citizens who are high self-monitors and relatively ineffective on citizens who are low self-monitors. Further research might test these predictions.

In this chapter we have explored several ways that the three facets of our account of why people vote could be incorporated into GOTV strategy. The three facets are summarized in Table 5.1, with relevant areas of behavioral research for each facet, as well as the major GOTV tactics that follow from each area of research. We believe that our approach has both practical and theoretical value. Practically, most of the behavioral principles that we cite have not yet been widely recognized by practitioners and policy makers. Although some current best practices implicitly leverage some of these theories, innovations in GOTV strategy are not systematically guided by these insights. We hope that providing a limited set of scientifically grounded behavioral principles to voter mobilization experts will help them devise more effective GOTV methods. At the same time, we hope that policy makers who are interested in increasing voter participation find this framework useful. For example, policy makers may publish the names of those who do and do not vote as regular practice or leverage other public services around election time to facilitate vote plan-making, or incorporate social-norm information into ballot guidebooks that are mailed to citizens before elections. It is worth noting that whereas more personal modes of GOTV contact tend to be more costly than less personal modes, more effective GOTV messages are generally no more costly than less effective messages, suggesting that using some

Table 5.1 Implications of voting as dynamic social behavior

Implication	Behavioral research	Recommended GOTV tactic
Dynamic: voting affected by events before and after decision	Self-prediction and commitment	Elicit vote intention (especially public commitments)
	Implementation intentions	Ask how, when, where, about voting
	Social pressure and accountability	Make voting records publicly accessible
Social: voting influenced by affiliative and belonging needs	Social identity	Emphasize benefits to favored others (in-group members)
	Descriptive social norms	Emphasize high expected turnout
Expression: voting as an expression of identity	Self-perception, social identity	Label, or make salient, a (social) identity that encourages voting
	Cognitive dissonance	Facilitate small steps, foot-in-the-door
	Correspondence bias	Facilitate picture of oneself voting from the third-person perspective

of the above behavioral principles in GOTV efforts could result in costless increases in impact.

Theoretically, by testing the behavioral principles discussed in this section in the GOTV context we might better understand the roles that each of these variables play in citizens' decisions to participate in democratic elections. In addition to expanding our understanding of why people vote, testing these behavioral principles in the GOTV context can also provide opportunities to learn more about the moderators and mediators of behavioral phenomena and also provide insight into how these phenomena interact. To cite one example, the discovery by Gerber and Rogers (2009) that some subgroups appear to be unaffected by descriptive social norms whereas others are very much affected by them provides not only a novel practical insight but also a novel theoretical insight into the study of descriptive social norms and suggests a direction for continuing theoretical research.

Communications around voter mobilization are just one type of political communication. Others include policy communications, campaign persuasion, candidate debates, and fund-raising communications, to name a few. The behavioral insights described in this chapter—as well as others not described herein—probably also apply to these areas. This exploration of why people vote illustrates some potential synergy between behavioral research and field experimentation. A more realistic behavioral model of individual behavior can generate new approaches for more effectively influencing voters. Moreover, field research that systematically investigates these behavioral principles can generate new insights for enriching theoretical models of human behavior. For these reasons, we foresee behavioral approaches playing an increasingly prominent role in research on political communications and on best practices among political professionals.

Notes

1. There are several reasons why enhancing voter turnout is a socially desirable objective. First, because elected officials have an incentive to represent the interests of the individuals they expect to vote in future elections, maximizing participation results in broadening the constituency that holds government accountable and to which government must be responsive. Second, when people vote they tend to see themselves as more civically engaged and thus may be more likely to engage in other civic activities (Finkel, 1985; Gerber, Green, and Shachar, 2003). Third, higher turnout increases the perceived legitimacy of an elected government, which increases the perceived legitimacy of the laws it enforces. Additionally, stimulating turnout in a given election encourages habitual voting behavior; inducing voting in the present election increases the likelihood of continued voting in the future (Gerber, Green and Shachar, 2003). To the extent that we accept that greater turnout is socially desirable, this means that successful GOTV has beneficial intermediate-term consequences in addition to immediate ones.

2. In light of more recent studies demonstrating a large effect for communications with striking messages (see the sections "Dynamic," "Social," and "Identity"), this conclusion is softened somewhat in their subsequent reviews of the literature to "subtle variations" have little effect (Green and Gerber, 2008, p. 70).

3. Estimating the cost per net vote generated requires estimating the cost per contact. Since these estimates vary widely, there is no universal answer to the question of how much it costs to generate a single new vote. For more information on this topic, see Green and Gerber (2008).

4. That said, if one makes a first, extreme request that is rejected and is immediately followed by a second, smaller request, this can increase compliance to the second request because the target may feel compelled to reciprocate the concession made by the requester (e.g., Cialdini et al., 1975).

5. This is consistent with a recent meta-analysis of fourteen GOTV canvassing experiments. It found that mobilization efforts affect citizens whose past vote history suggests they were on the cusp of whether or not to turn out in a given election; that is, they have a moderate probability of voting in a given election (Arceneaux and Nickerson, 2009).

References

Allen, C. T. (1982). Self-perception based strategies for stimulating energy conservation. *Journal of Consumer Research*, *8* (4), 381–390.

Arceneaux, K. (2005). Using cluster randomized field experiments to study voting behavior. The science of voter mobilization. *Annals of the American Academy of Political and Social Science*, *601*, 169–179.

Arceneaux, K., and Nickerson, D. W. (2006). *Even if you have nothing nice to say, go ahead and say it: Two field experiments testing negative campaign tactics.* Paper presented at the 2005 meeting of the American Political Science Association, September 1–4, Washington, DC.

———. (2009). Who is mobilized to vote? A re-analysis of eleven field experiments. *American Journal of Political Science*, *53*(1), 1–16.

Baumeister, R. F., and Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, *117*(3), 497–529.

Belk, R. W. (1988). Property, persons, and extended sense of self. In L. F. Alwitt (Ed.), *Proceedings of the Division of Consumer Psychology, American Psychological Association 1987 Annual Convention* (pp. 28–33). Washington, DC.: American Psychological Association.

Bem, D. J. (1972). Self-perception theory. *Advances in experimental Social Psychology*, *6*, 1–62.

Bennion, E. A. (2005). Caught in the ground wars: Mobilizing voters during a competitive congressional campaign. *Annals of the American Academy of Political and Social Science*, *601*, 123–141.

Blais, A. 2000. *To vote or not to vote: The merits and limits of rational choice theory.* Pittsburgh, PA: University of Pittsburgh Press.

Burger, J. M. (1999). The foot-in-the-door compliance procedure: A multiple-process analysis and review. *Personality and Social Psychology Review*, *3* (4), 303–325.

Burger, J. M., Messian, N., Patel, S., del Prado, A., and Anderson, C. (2004). What a coincidence! The effects of incidental similarity on compliance. *Personality and Social Psychology Bulletin*, *30*(1), 35–43.

Camerer, C. F., (2003). *Behavioral game theory: Experiments in strategic interaction.* Princeton, NJ: Princeton University Press; New York: Russell Sage Foundation.

Cann, A., Sherman, S. J., and Elkes, R. (1975). Effects of initial request size and timing of a second request on compliance: The foot in the door and the door in the face. *Journal of Personality and Social Psychology*, *32*(5), 774–782.

Chartrand, T., Pinckert, S., and Burger, J. M. (1999). When manipulation backfires: The effects of time delay and requester on foot-in-the-door technique. *Journal of Applied Social Psychology*, *29* (1), 211–221.

Cialdini, R. B. (2003). Crafting normative messages to protect the environment. *Current Directions in Psychological Science*, *12*(4), 105–109.

Cialdini, R., Demaine, L. J., Sagarin, B. J., Barrett, D. W., Rhoads, K. and Winters, P. L. (2006). Activating and aligning social norms for persuasive impact. *Social Influence*, *1*, 3–15.

Cialdini, R. B., Reno, R. R., and Kallgren, C. A. (1990). A focus theory of normative conduct—Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, *58*(6), 1015–1026.

Cialdini, R. B., Vincent, J. E., Lewis, S. K., Caralan, J., Wheeler, D., and Draby, B. L. (1975). Reciprocal concessions procedure for inducing compliance: The door-in-the-face technique. *Journal of Personality and Social Psychology*, *31*, 206–215.

Coate, S., and Conlin, M. (2004). A group rule-utilitarian approach to voter turnout: Theory and evidence. *American Economic Review*, *94*(5), 1476–1504.

Dale, A., and Strauss, A. (2007). *Text messaging as a youth mobilization tool: An experiment with a post-treatment survey.* Paper presented at the annual meeting of the Midwest Political Science Association, Chicago, IL.

Deutsch, M., and Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *Journal of Abnormal Psychology*, *51*(3), 629–636.

Dholakia, U. M., and Bagozzi, R. P. (2003). As time goes by: How goal and implementation intentions influence enactment of short-fuse behaviors. *Journal of Applied Social Psychology*, *33*, 889–922.

Downs, A. (1957). *An economic theory of democracy.* New York: Harper and Row.

Edlin, A., Gelman, A., and Kaplan, N. (2007). Voting as a rational choice: Why and how people vote to improve the well-being of others. *Rationality and Society*, *19*, 293–314.

Feddersen, T., and Sandroni, A. (2006). A theory of participation in elections. *American Economic Review*, *96* (4), 1271–1282.

Festinger, L. (1964). *Conflict, decision, and dissonance.* Stanford, CA: Stanford University Press.

Festinger, L., and Carlsmith, J. M. (1959). The cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology*, *58*, 203–210.

Finkel, S. E. (1985). Reciprocal effects of participation and political efficacy: A panel analysis. *American Journal of Political Science*, *29*(4), 891–913.

Fiorina, M. P. (1974). The paradox of not voting: A decision theoretic analysis. *American Political Science Review*, *68*(2), 525–536.

Fournier, R., Sosnik, D. B., and Dowd, M. (2006). *Applebee's America: How successful political, business, and religious leaders connect with the new American community.* New York: Simon and Shuster.

Fournier, S. (1998). Consumers and their brands: Developing relationship theory in consumer research. *Journal of Consumer Research*, *24*(4), 343–373.

Forsythe, R., Horowitz, J. L., Savin, N. E., and Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, 6, 347–69.

Fowler, J. H., and Kam, C. D. (2006). Patience as a political virtue: Delayed gratification and turnout. *Political Behavior*, *28*(2), 113–128.

———. (2007). Beyond the self: Social identity, altruism, and political participation. *Journal of Politics*, *69*(3), 813.

Franklin, D. P., and Grier, E. E. (1997). Effects of motor voter legislation. *American Politics Quarterly*, *25*(1), 104.

Freedman, J. L., and Fraser, S. C. (1966). Compliance without pressure—Foot-in-the-door technique. *Journal of Personality and Social Psychology*, *4*(2), 195–202.

Gangestad, S. W., and Snyder, M. (2000). Self-monitoring: Appraisal and reappraisal. *Psychological Bulletin*, *126*(4), 530–555.

Gerber, A. S., Gimpel J. G., Green, D. P., and Shaw D. R. (2006). *The influence of television and radio advertising on candidate evaluations: Results from a large scale randomized experiment.* Paper presented at the

Standard University Methods of Analysis Program in the Social Sciences, Palo Alto, CA.

Gerber, A. S., and Green, D. P. (2000a). The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment. *American Political Science Review*, *94*(3), 653–663.

———. (2000b). The effect of a nonpartisan get-out-the-vote drive: An experimental study of leafleting. *Journal of Politics*, *62*, 846–857.

———. (2001). Do phone calls increase voter turnout? *Public Opinion Quarterly*, *65*(1), 75–85.

———. (2005). Correction to Gerber and Green (2000), Replication of disputed findings, and reply to Imai (2005). *American Political Science Review*, *99*(2), 301–313.

Gerber, A. S., Green D. P., and Larimer C. W. (2008). Social pressure and voter turnout: Evidence from a large-scale field experiment. *American Political Science Review*, *102*, 33–48.

Gerber, A. S., Green, D. P., and Shachar, R. (2003). Voting may be habit-forming: Evidence from a randomized field experiment. *American Journal of Political Science*, *47*(3), 540.

Gerber, A. S., and Rogers, T. (2009). Descriptive social norms and motivation to vote: Everybody's voting and so should you. *Journal of Politics*, *71* (1), 1–14.

Gertner, J. (2004, February 15). The very, very personal is the political. *New York Times Magazine*, Section 6, p. 43–43. Retrieved from http://www.nytimes.com/2004/02/15/magazine/15VOTERS.html?page wanted=all

Gilbert, D. T., and Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin*, *117*, 21–38.

Goldstein, N. J., Cialdini, R. B., and Griskevicius, V. (2008). A room with a viewpoint: Using norms to motivate environmental conservation hotels. *Journal of Consumer Research*, *35*, 472–481.

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist*, *54*, 493–503.

Gollwitzer, P. M., Bayer, U., and McCulloch, K. (2005). The control of the unwanted. In R. Hassin, J. Uleman, and J. A. Bargh (Eds.), *The new unconscious* (pp. 485–515). Oxford: Oxford University Press.

Gollwitzer, P. M., and Sheeran, P. (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 38, pp. 69–119). San Diego, CA: Academic Press.

Green, D. P. and Gerber, A. S. (2004). *Get out the vote!* Washington, DC: Brookings Institution Press.

———. (2008). *Get out the vote!* (Rev. ed.).Washington, DC: Brookings Institution Press.

Green, D. P., Gerber, A. S., and Nickerson, D. W. (2003). Getting out the vote in local elections: Results from six door-to-door canvassing experiments. *Journal of Politics*, *65*(4), 1083–1096.

Green, D. P., and Vavreck, L. (2006). *Assessing the turnout effects of Rock the Vote's 2004 television commercials: A randomized field experiment.* Paper presented at the Annual Meeting of the Midwest Political Science Association, Chicago, IL.

Greenwald, A. G., Carnot, C. G., Beach, R., and Young, B. (1987). Increasing voting-behavior by asking people if they expect to vote. *Journal of Applied Psychology*, *72*(2), 315–318.

Greenwald, A. G., Klinger, M. R., Vande Kamp, M. E., and Kerr, K. L. (1988). *The self-prophecy effect: Increasing voter turnout by vanity-assisted consciousness raising.* Unpublished manuscript, University of Washington, Seattle.

Hansen, R. A., and Robinson, L. M. (1980). Testing the effectiveness of alternative foot-in-the-door manipulations. *Journal of Marketing Research*, *17*(3), 359–364.

Iyengar, S. (2002). *Experimental designs for political communication research: From shopping malls to the Internet.* Unpublished manuscript, Stanford University.

Jankowski, R. (2002). Buying a lottery ticket to help the poor: Altruism versus self-interest in the decision to vote. *Rationality and Society*, *14*(1), 55–77.

Jones, E. E., and Nisbett, R. E. (1971). The actor and the observer: Divergent perceptions of the causes of behavior. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, and B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 79–94). New York: General Learning Press.

Kerr, N. L., and MacCoun, R. J. (1985). The effects of jury size and polling method on the process and product of jury deliberation. *Journal of Personality and Social Psychology*, *8*, 319–323.

Kiecolt-Glaser, J. K., Garner, W., Speicher, C., Penn, G. M., Holliday, J., and Glaser, R. (1984). Psychosocial modifiers of immunocompetence in medical students. *Psychosomatic Medicine*, *46*, 7–14.

Lerner, J. S., and Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*, 255–275.

Libby, L. K., Eibach, R. P., and Gilovich, T. (2005). Here's looking at me: The effect of memory perspective on assessments of personal change. *Journal of Personality and Social Psychology*, *88*(1), 50.

Light, P. C. (1999). *The true size of government.* Washington, DC: The Brookings Institution Press.

Lijphart, A. (1997). Unequal participation: Democracy's unresolved dilemma. Presidential address, American Political Science Association, 1996. *American Political Science Review*, *91*(1), 1–13.

Lippke, S., and Ziegelmann, J. P. (2002). *Self-regulation and exercise: a study on stages of change and successful ageing*. Unpublished manuscript, Free University of Berlin, Germany.

Lynch, J. J. (1979). *The broken heart: The medical consequences of loneliness*. New York: Basic Books.

Lyubomirsky, S., Sheldon, K. M., and Schkade, D. (2005). Pursuing happiness: The architecture of sustainable change. *Review of General Psychology*, *9*(2), 111–131.

Mann, C. B. (2009). *Resolving the mobilization conundrum of message and personal contact: Voter registration, vote by mail, and election day voting field experiments*. Unpublished manuscript, Yale University.

Mattern, J. D. and Neighbors, C. (2004). Social norms campaigns: Examining the relationship between changes in perceived norms and changes in drinking levels. *Journal of Studies on Alcohol*, *65*, 489–493.

McNulty, J. E. (2005). Phone-based GOTV—What's on the line? Field experiments with varied partisan components, 2002–2003. *Annals of the American Academy of Political and Social Science*, *601*, 41–65.

Michelson, M. R. (2003). Getting out the Latino vote: How door-to-door canvassing influences voter turnout in rural central California. *Political Behavior*, *25*(3), 247–263.

———. (2005). Meeting the challenge of Latino voter mobilization. *Annals of the American Academy of Political and Social Science*, *601*, 85–101.

Michelson, M. R., McConnell, M. A., and Bedolla, L. G. (2009). *Heeding the call: The effect of targeted phone-banks on voter turnout*. Manuscript submitted for publication.

Middleton, J. A., and Green, D. P. (2008). Do community-based voter mobilization campaigns work even in battleground states? Evaluating the effectiveness of MoveOn's 2004 outreach campaign. *Quarterly Journal of Political Science*, *3*, 63–82.

Miller, R. L., Brickman, P., and Bolen, D. (1975). Attribution versus persuasion as a means for modifying behavior. *Journal of Personality and Social Psychology*, *31* (3), 430–441.

Milne, S., Orbell, S., and Sheeran, P. (2002). Combining motivational and volitional interventions to promote exercise participation: Protection motivation theory and implementation intentions. *British Journal of Health Psychology*, *7*(2), 163.

Morwitz, V. G., Johnson, E., and Schmittlein, D. (1993). Does measuring intent change behavior? *Journal of Consumer Research*, *20*(1), 46–61.

Murray, G. R., and Matland, R. E. (2005). *Increasing voter turnout in the Hispanic community: A field experiment on the effects of canvassing, leafleting, telephone calls, and direct mail*. Paper presented at the Midwest Political Science Association 63rd Annual National Conference.

Myers, D. (1992). *The pursuit of happiness*. New York: Morrow.

Nickerson, D. W. (2005). Scalable protocols offer efficient design for field experiments. *Political Analysis*, *13*(3), 233–252.

———. (2006a). *Demobilized by e-mobilization: Evidence from thirteen field experiments*. Unpublished Manuscript. Department of Political Science, University of Notre Dame.

———. (2006b). *Forget me not? The importance of timing in voter mobilization*. Paper presented at the annual meeting of the American Political Science Association, Philadelphia, PA.

———. (2006c). Hunting the elusive young voter. *Journal of Political Marketing*, *5*(3), 47–69.

———. (2006d). Volunteer phone calls can increase turnout. *American Politics Research*, *34*(3), 271–292.

———. (2007). Quality is job one: Professional and volunteer voter mobilization calls. *American Journal of Political Science*, *51*(2), 269–282.

Nickerson, D. W., Friedrichs, R. D., and King, D. C. (2006). Partisan mobilization campaigns in the field: Results from a statewide turnout experiment in Michigan. *Political Research Quarterly*, *59*(1), 85–97.

Nickerson, D. W., and Rogers, T. (2010). Do you have a voting plan? *Psychological Science*, *21*(2), 194–199. doi:10.1177/0956797609359326

Panagopoulos, C., and Green, D. P. (2006), *The impact of radio advertisements on voter turnout and electoral competition*. Paper presented at the Annual Meeting of the Midwest Political Science Association, Chicago, IL.

Posner, R. A., and Rasmusen, E. B. (1999). Creating and enforcing norms, with special reference to sanctions. *International Review of Law and Economics*, *19*(3), 369–382.

Reams, M. A., and Ray, B. H. (1993). The effects of 3 prompting methods on recycling participation rates—A field study. *Journal of Environmental Systems*, *22*(4), 371–379.

Reno, R. R., Cialdini, R. B., and Kallgren, C. A. (1993). The transsituational influence of social norms. *Journal of Personality and Social Psychology*, *64*(1), 104.

Riker, W. H., and Ordeshook, P. C. (1968). Theory of calculus of voting. *American Political Science Review*, *62*(1), 25–42.

Rind, B., and Benjamin, D. (1994). Effects of public image concerns and self-image on compliance. *Journal of Social Psychology*, *134*, 19–25.

Rogers, T. (2005). [Experts in voter mobilization lack strong intuition on effectiveness of descriptive social norms.] Unpublished data.

Schlenker, B. R., Dlugolecki, D. W., and Doherty, K. (1994). The impact of self-presentations on self-appraisals and behavior—The power of public com-

mitment. *Personality and Social Psychology Bulletin*, *20*(1), 20–33.

Sheeran, P., and Orbell, S. (1999). Implementation intentions and repeated behaviour: Enhancing the predictive validity of the theory of planned behaviour. *European Journal of Social Psychology*, *29*, 349–369.

Sherman, S. J. (1980). On the self-erasing nature of errors of prediction. *Journal of Personality and Social Psychology*, *39*, 211–221.

Smith, J. K., Gerber, A. S., and Orlich, A. (2003). Self-prophecy effects and voter turnout: An experimental replication. *Political Psychology*, *24*(3), 593–604.

Snyder, M. (1974). Self–monitoring of expressive behavior. *Journal of Personality and Social Psychology, 30(4)*, 526–537.

Snyder, M., and Cunningham, M. R. (1975). To comply or not comply: Testing the self-perception explanation of the foot-in-the-door phenomenon. *Journal of Personality and Social Psychology*, *31*, 64–67.

Stollwerk, A. F. (2006). *Does e-mail affect voter turnout? An experimental study of the New York City 2005 election*. Unpublished manuscript, Institution for Social and Policy Studies, Yale University.

Swann, W. B., Jr., and Read, S. J. (1981). Self-verification processes: How we sustain our self-conceptions. *Journal of Experimental Social Psychology*, *17*, 351–372.

Tajfel, H. (1982). *Social identity and intergroup behavior*. Cambridge: Cambridge University Press.

Tullock, G. (1968). Towards a mathematics of politics. Ann Arbor, MI: University of Michigan Press.

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., and Wetherell, M. (1987). *Rediscovering the social group: A self-categorization theory*. Oxford: Basil Blackwell.

Tybout, A. M., and Yalch, R. F. (1980). The effect of experience—A matter of salience. *Journal of Consumer Research*, *6*(4), 406–413.

CHAPTER 6

Perspectives on Disagreement and Dispute Resolution

Lessons from the Lab and the Real World

LEE ROSS

Public policy generally results from discussions and negotiation between parties who disagree—discussions and negotiations that can either exacerbate or attenuate ill will. This chapter explores some of the processes that add hostility and distrust to policy disagreements, sentiments that make conflicts between antagonists more difficult to resolve. But other barriers make their influence felt as well (Mnookin and Ross, 1995). Deception, intransigence, and other tactics often impede the achievement of efficient agreements; political considerations and agency problems can also play a role. The barriers to be discussed in detail here, however, are ones that arise from *psychological* processes and biases.

The initial part of the chapter begins with a brief discussion of the role of subjective interpretation in leading individuals and groups to see things differently, to favor different policies, or to work toward different changes in the status quo. But the major focus of this section of the chapter is on an epistemic stance, termed "naive realism" (Ross and Ward, 1995, 1996; see also earlier work by Griffin and Ross, 1991; Ichheiser 1949, 1970), that is, the conviction that one sees and considers issues and events in an objective and essentially unmediated fashion. This conviction adds rancor to disagreement insofar as it leads opposing partisans to feel that other reasonable people ought to share their views, and hence that those who view things *differently* are *un*reasonable—in particular, that those on the "other side" are displaying self-serving biases and other sources of distortion to which they and those sharing their views are relatively immune.

Naive convictions about the status of one's own versus other's perceptions and judgments make productive discussions and negotiations more difficult, but various cognitive and motivational processes, most of which are familiar to social psychologists and to researchers in the judgment and decision-making tradition, create additional barriers. These barriers, especially the *reactive devaluation* of proposals put on the table by the other side, are discussed in the second part of this chapter.

The final part of the chapter deals with strategies and tactics for *overcoming* barriers to agreement. Theory and research of the sort that are normally reported in academic journals are reviewed, but I also will take this opportunity to reflect on insights offered by my own "real world" experiences and those of my colleagues at the Stanford Center on International Conflict and Negotiation (SCICN). These experiences, gained in efforts at second-track diplomacy, intercommunity dialogue, and other "public peace processes" (Kelman 2001; Saunders, 1999) in Northern Ireland and the Middle East, have served both to underscore the importance of the theory and research reviewed here and to offer directions for further investigation.

Divergent Views and Perceptions of Bias

Conflict arises when parties disagree about a future course of action. Such disagreement reflects differences in interests, values, priorities, and expectations. But not all disputes lead to conflict. Moreover, some conflicts are harsher, more intractable, and costlier than others. To appreciate the relationship between disagreement and enmity, it is important to note that disputing parties do more than simply calculate their interests and adopt strategies for forwarding those interests. They observe and interpret each other's words and deeds. They make predictions about what ought to and what will happen, and they have affective reactions to what they observe, hear, and anticipate.

One of social psychology's most enduring contributions to the understanding of disagreement has been to highlight the importance of subjective

interpretation. Long ago, Solomon Asch (1952) cautioned us that differences in judgment might reflect differences not in values or preferences, but rather in the way the "objects of judgments" are being perceived or construed by the relevant individuals. In a paper that helped to launch the cognitive revolution in psychology, Jerome Bruner (1957) observed that people go "beyond the information given." They fill in details of context and content, they infer linkages between events, and they use their existing dynamic scripts or schemes or adopt new ones to give events coherence and meaning. (See Nisbett and Ross, 1980; also, Fiske and Taylor, 1984, 2008).

Successful politicians have long recognized that success in public policy debates depends in part on controlling the way in which the relevant issues are construed. In the midst of the Depression, Franklin Roosevelt and his New Deal lieutenants anticipated that the proposed new program of intergeneration income transfer, which we now know today as the social security system, would be characterized by opponents as an ill-advised venture in welfare or, worse, socialism. Accordingly, the plan to deduct a portion of each worker's income to fund the new system was portrayed as a kind of personal savings or pension plan. The image put forward was one of a steadily accumulating nest egg to be tapped in one's golden years, with an accompanying insurance policy to provide continuing income if unanticipated misfortune struck. Not coincidentally, there was no explicit acknowledgment that the first generation of beneficiaries of this plan would receive much more from the plan than they contributed; nor was it suggested that subsequent generations of workers might be obliged to pay more and receive a less generous return. The truth, of course, was that there were no gradually accumulating individual accounts. There was only the government's promise to meet financial obligations as they arose.

In the years following, the debate about the social security system has become increasingly heated. Critics complain that having today's worker support retirees who paid relatively little constitutes a kind of pyramid scheme, one that is bound to collapse when not enough new "suckers" can be persuaded to pour in fresh money. A more objective characterization than that offered by either the New Deal proponents or the later conservative critics, of course, is that the federal government is simply taxing current wage earners and providing benefits to retired or disabled workers (and, after the workers' deaths, to their dependents) in fulfillment of the same kind of social contract to be found in virtually every industrialized country in the world. The system is in little danger of "collapsing" unless the government itself goes bankrupt; although, over the long haul, total government expenditures, of all sorts, including social security payouts, will have to be balanced (or nearly so) by the total government revenues.

The political battle to manipulate construals and thereby win support or marshal opposition to particular public policies goes on constantly. Depending on the views and interests of those controlling the media, we hear references to "illegal aliens" versus "undocumented workers," to "terrorists" versus "insurgents," to "democracy building" versus "putting of our troops in the middle of a civil war." As George Orwell warned so chillingly in *1984*, those who have the capacity to control language and media, and thus to control the schemas we utilize in considering policies and events, enjoy the power to control political attitudes and behavior.

Three decades ago, investigators in cognitive and social psychology showed that increasing the saliency of particular schemas and knowledge structures can influence the respondents' behavioral choices (Gilovich, 1981; Higgins, Rholes, and Jones, 1977). Ever more sophisticated and compelling demonstrations of "priming effects" on social and political preferences and of overt behavior are provided today in a number of laboratories. For example, Bryan et al. (2009) showed that support for liberal versus conservative policies among college students could be manipulated through a prior priming task that called for them to describe the role either that "good fortune and the help of others" or "hard work and good decision making" had played in helping them win admission to Stanford University.

Rather than digress here to review the empirical demonstrations that social psychologists in successive generations have offered in hoisting the subjectivist banner, let me merely offer two assertions: The first and more obvious assertion is simply that differences in subjective interpretation *matter*, that they have a profound impact in the conduct of everyday social affairs. The second and less obvious assertion is that social perceivers characteristically make *insufficient allowance* for such impact in the inferences and predictions they make about other individuals. In the particular case of opposing partisans who look at the facts and history of events relevant to their conflict, both are apt to find additional support for their preexisting views. As a result, the two sides are apt to become more, instead of less, polarized in their sentiments and beliefs as a consequence of their involvement in the debate and to respond negatively when they hear the other side characterize the "facts" in question.

Lord, Ross, and Lepper (1979) explored the consequences of such *biased assimilation* in the responses of death-penalty proponents and opponents to the purported mixed results of a pair of studies

that employed differing methodologies (i.e., contrasting homicide rates in adjacent death-penalty and no-death-penalty states and contrasting rates before and after changes in statutes permitting or ending executions). Although the investigators employed a carefully balanced design that matched positive and negative findings equally often with each of the different methodologies, both sides proceeded to accept uncritically the results of the study supporting their position and to identify obvious flaws in the study opposing their position—and thus, as predicted, to become further polarized in their views as they assimilated the relevant findings.

While participants in the Lord, Ross, and Lepper study were not asked to make attributions about the assessments of the relevant research by those on the other side of the capital punishment debate, it is not difficult to imagine that it would be less than charitable. One can expect similar assimilation biases to operate, and even more negative attributions to be offered, in the context of other conflictual issues in which not only the interpretation of facts, but the facts themselves are in dispute. Partisans on both sides can be expected to characterize those on the other side—especially those who claim to derive support for their position from an "objective" reading of the historical facts and other evidence—as either dishonest or deluded. Indeed, this scenario for biased assimilation, imputations of bias, and other hostile attributions in the context of ongoing conflict anticipates the discussion of naive realism to follow.

Convictions of the Naive Realist

When people, whether laypeople or sophisticated policy makers, confront political issues and actors, they are apt to do so with a confidence rooted in the conviction that a one-to-one relationship exists between their personal *perception* of the external objects and events and the "real" nature of the objects and events themselves. Expressed in first-person terms, my basic conviction as a naive realist is that I see entities and events as they are in objective reality, that my basic perceptions and the attitudes, preferences, sentiments, and priorities that arise from those perceptions reflect a relatively dispassionate, unbiased, and essentially "unmediated," or "bottom-up," rather than "top-down" apprehension of the information or evidence at hand.

From this conviction, it follows, therefore, that other rational social perceivers generally will, and in fact *should*, share not only my perceptions but also the opinions, judgments, priorities, and feelings that follow from those perceptions—provided that they have access to the same information as I do and provided that they too have processed that information in a thoughtful and open-minded fashion. There is by now a large body of evidence for such a *false consensus* effect (Marks and Miller, 1987; Ross, Greene and House, 1977); that is, people who make a given personal choice or hold a particular view tend to see that choice or view as more common, and less revealing of distinguishing personal attributes, than do people who make the opposite choice or hold the opposite view. Moreover, as Gilovich (1990) demonstrated in a clever series of studies, the *magnitude* of this effect proves to be a function of the degree to which the object of judgment to which research participants are responding is one that offers latitude for different interpretations or "construals."

It also follows that since I see things as they are, the failure of a given individual or group to share my views must reflect some deficiency in *them*, or more specifically, some deficiency in the process by which they have arrived at their wrong views. One possibility is that they are unable or unwilling to proceed from objective evidence to reasonable conclusions; in which case, their opinions can and should be ignored. A second possibility is that they have not yet been exposed to the "real facts" or had their attention focused on the "real issue"; in which case, provided that they are reasonable and open-minded, once I expose them to the real facts and give them the proper perspective, they will see the light and come to agree with me. In this regard it is instructive to consider the following newspaper account of a meeting between President George W. Bush and a distinguished group of former cabinet members and foreign policy experts:

> The president joined Gen. George Casey, the top American Commander in Iraq, and Zalmay Khalilzad, the U.S. ambassador in Baghdad, to give a detailed briefing on Iraq to more than a dozen foreign policy leaders from previous administrations, split nearly evenly between Democrat and Republican [including Colin Powell, Madeleine Albright, George Shultz, Robert McNamara, James Baker, Melvin Laird, William Perry, Lawrence Eagleburger, William Cohen, and Harold Brown]. . . . The White House's hope was that the prominent figures—many of whom have publicly opposed Bush on Iraq—would be persuaded by the president's argument that he has what he called a "dual-track strategy for victory," and they would then spread the word. (*USA Today*, January 5, 2006)

If we take this account at face value, the president, even in the face of events that were challenging the wisdom of his strategy and tactics, was not considering the possibility that the perceptions and

analyses prompting that strategy and those tactics were incorrect. Instead, his objective was to have these distinguished veterans of statecraft see things as he did—that is, "correctly"—so that they would get on board and help him answer his critics. There is a considerable literature on the use and misuse of advisors, and in particular on the phenomenon of *groupthink*, whereby bad decisions are made because the benefits of divergent perceptions and judgments, and of airing doubts about potential pitfalls and ways to reduce risks or costs, are denied to the decision maker (Janis, 1972; Janis and Mann 1977). However the tendency to treat dialogue opportunities as an opportunity to influence, rather than learn or open oneself to the influence of others, is by no means unique to overly confident leaders. In our experience at SCICN, well-meaning moderates who come to dialogue groups and other citizen-based processes designed to help moderates on the two sides find common ground, do so intending to listen attentively to the views of the other side, but also to use the opportunity to explain how things "really" are. They hope and expect that counterparts on the other side who come in good faith can thereby be persuaded to change their views. When that hope is not realized, as is usually the case, the risk is that they will leave persuaded that their counterparts have not come in good faith or that they continue to be deluded.

Assuming that those on the other side of an issue care and are capable of following an argument and that they are not uninformed, a third possible explanation for others' failure to see things my way, is that *they* (unlike me and those who agree with me) are not being objective. The inference is that *their* judgments and the positions they are advocating have been distorted by self-interest, by some pernicious ideology, or some other source of bias (to which I, fortunately, am immune). It is this explanation for disagreement that has been the object of most intense investigation in our own program of research, some of which I describe below.

The inference that others are seeing matters in an inaccurate, even systematically biased, manner does not immediately produce enmity. On the contrary, it may well lead the naive realist to assume that rational open-minded discourse, in which information and cogent arguments are freely exchanged, will lead to agreement (or at least to a marked narrowing of disagreement). Such optimism, however, generally proves to be short-lived. While the experience of the dialogue participants may be positive in many respects, neither side generally yields much to the other side's attempts at enlightenment. The conclusion reached by individuals on both sides of the issue, especially when it is clear that those on the other side are not lacking in interest or intellectual capacity, is that the ability of those on the other side to proceed from facts and evidence to conclusions is being distorted by some combination of self-interest, defensiveness, mistaken ideology, or other cognitive or emotional biases.

Attributions involving individual or collective self interest are, in fact, apt to be buttressed by observation and analysis. There generally *is* a correlation between beliefs held or policies advocated and the individual or collective self-interest of the relevant advocates. Naive realists thus rarely find it difficult to detect that linkage in the policies favored by those on the other side of the table. What they generally lack is recognition that a similar correlation exists between their own views and their own self-interests.

Before reviewing research on attributions of bias in self versus others, there is one caveat to be added to the present account of the naive realist's claims of pure objectivity. People who share a particular group identity, formative experience, or basis for self-interest often acknowledge that their views are influenced, or "informed," by that source of identity. However, they claim that *their* special in-group identity is a source of *insight* and appropriate sentiments, whereas the in-group identity of *others*, particularly others holding different views, is a source, not of enlightenment, but of *bias*. This phenomenon was demonstrated in a simple study conducted at Cornell University in which Caucasian and non-Caucasian students were asked their views about affirmative action, and varsity and intramural athletes were asked about their views on the use of university athletic facilities—and then in each case, the participants were asked to assess the "enlightening versus distorting" influence of their own status versus that of the opposing interest group on those views (Ehrlinger, Gilovich, and Ross, 2005). As predicted, a group's own identity was consistently seen more as a source of enlightenment and insight, whereas the other group's identity was consistently seen as a source of self-interested distortion.

Biased Perceptions and Perceptions of Bias on the Part of the "Other" Side

Emily Pronin and her colleagues (including the present author) have explored the tendency for people to impute bias to people holding views different from their own. This work, which dealt specifically with perceived bias versus objectivity in judgments about political issues that typically divide Americans, is described in considerable detail in Pronin's chapter in this volume. The central finding in that work is that people see themselves as less subject to a large number of specific biases (including wishful thinking,

self-serving biases in interpreting data, and a variety of other psychological failings that both produce bad decisions and exacerbate conflict) than are other individuals. Furthermore, the tendency for people to see others as more susceptible to such biases than they are appears to be a linear function of the perceived discrepancy between their own views and the views of those "others" (Pronin, Gilovich, and Ross, 2004).

In an earlier and somewhat related line of work, investigators of the so-called third-person effect report that people exposed to persuasive communications, or "propaganda," in the mass media believe that it will have a greater effect on others than on themselves (Davison, 1983; see also Paul, Salwen, and Dupagne, 2000). Such expectations and the fears they engender increase support for restrictions on mass media and for other measures that would restrict free speech. Of course, those who embrace a particular message or regard a piece of propaganda or disinformation as simple truth express no such fears—they worry that others will not "see the light" and act accordingly.

The Hostile Media (and Mediator) Effect

Naive realism can be a source of negative attributions not only about adversaries but also about third parties who venture opinions about the relative merits of the claims made by the opposing sides. Ubiquitous claims of media bias by those both on the left and the right in American politics are a case in point. Both groups claim that the other side has disproportionate access to the airwaves and editorial pages of major newspapers, and both think that the other side's claims to the contrary are cynical and dishonest. Such accusations and counteraccusations can indeed be politically motivated. However, they are also a manifestation of naive realism; that is, opposing partisans in social policy debates are likely to share the honest conviction that the media *are* biased against them and too inclined to parrot the claims of the other side and not do justice to the arguments on their side. Partisans who see issues as black or white are bound to see any effort at balanced media coverage—coverage that emphasizes both black and white or that paints issues in hues of gray—as biased against their side.

This implication of naive realism was tested in a study by Vallone, Ross, and Lepper (1985). Capitalizing upon the passionately held differences in opinion that people hold about the Arab-Israeli conflict, the investigators presented pro-Israeli, pro-Arab, and "neutral" students with excerpts from then current television news coverage of the massacre of civilians in two South Lebanon Palestinian refugee camps. Whereas the best informed and most knowledgeable of these "neutrals" rated the broadcast samples as relatively unbiased on the issue of Israeli responsibility and complicity, the evaluations offered by the two groups of partisans were less charitable. On every measure pro-Arab and pro-Israeli viewers alike agreed strongly only that the other side had been favored by the media, that their own side had been treated unfairly, and that the relevant biases in reporting had reflected the ideologies and self-interests of those responsible for the program. (It is worth noting as an aside that skilled mediators recognize and take steps to avoid the relevant pitfalls—in particular, by not offering their own views about the accuracy of respective characterizations of history, about the validity of claims, or especially about the nature of a fair or just agreement. Instead, as I will discuss in more detail later in this chapter, they seek to frame the negotiation process in terms of finding common ground and arranging "efficient" trades of concession.)

There was also evidence (reminiscent of Hastorf and Cantril's famous 1954 account of the divergent perceptions of college football partisans from different schools in assessing the rough play of the two teams) that the two partisan groups, in a sense, "saw" different programs. Whereas viewers supportive of Israel claimed that a higher percentage of the specific facts and arguments presented had been anti-Israel than pro-Israeli, viewers hostile to Israel assessed that balance in opposite terms. Both sides, furthermore, believed that neutral viewers of the program would be swayed in a direction hostile to their side and favorable to the other side.

False Polarization

Conflict, misunderstandings, and misattribution, as noted earlier, can result when individuals or groups fail to recognize that they have construed issues or events differently and thus essentially have responded to different objects of judgment. But naive realism can also play a role in leading opposing partisans to *over*estimate the divide. Once it becomes clear that others do not share one's opinions and perspectives (especially when they persist in their erroneous views in the face of information and arguments that ought to have "enlightened" them), the likely attribution is one of bias and closed-mindedness. Naive realists feel that while their own "bottom-up" construals of issues and events reflect the richness, complexity, ambiguity, and even contradictions of objective reality, the construals of those who express and persist in views different from their own must be "top-down" products of ideology and self-interest. As such, the views of people on the other side (but to some degree even those on their own side as well) are expected be more extreme, more ideologically consistent, and freer of

nuance and ambivalence than their own views. In short, most people on both sides of an issue, as well as the majority in the "middle" who see merits and flaws in the arguments offered by both sides, are apt to overestimate political polarization and to be unduly pessimistic about the prospects of finding common ground.

Evidence for this phenomenon was provided in a pair of studies by Robinson et al. (1995) that compared partisan group members' actual differences in construal with their assumption about such differences. One study dealt with pro-choice versus pro-life views relevant to the ongoing abortion rights debates (e.g., what kind of abortion scenario and considerations are common versus uncommon; also, what positive consequences and what negative consequences would be likely to follow from a tightening of abortion restrictions, etc.). The second study dealt with liberal versus conservative construals of specific events in the racially charged Howard Beach incident in which a black teenager was fatally injured by an automobile while he was running away from a group of white pursuers (e.g., who had started and who had exacerbated the initial confrontation, what had been the intentions and motives of the various participants in the incidents, etc.).

As expected, both sides provided many instances of construal differences, but almost invariably the magnitude of such differences was overestimated rather than underestimated by the partisans. More specifically, the partisans overestimated the degree of ideological consistency that both sides—especially the other side, but to some extent their own side as well—would show in the assumptions and construals they brought to the relevant issues. What is more, individual partisans in both studies felt that their own views were less driven by ideology than those of other partisans. It is worth noting that nonpartisan or neutral respondents in the study showed the same tendency to overestimate the extremity and ideological consistency of the partisan groups' construals as the partisans did themselves. That is, partisans and nonpartisans alike significantly overestimated the "construal gap" between the two sides and, in a sense, underestimated the amount of common ground to be found in the assumptions, beliefs, and values shared by the relevant parties.

Informal interviews with students, incidentally, revealed a source of these misperceptions and overestimations that lay beyond naive realism. Students reported that they rarely acknowledged to others the degree of ambivalence in their political beliefs—not in talking to their ideological allies (lest their resoluteness come into doubt) and not in talking to their ideological adversaries (lest their concessions be exploited or misunderstood). In fact, most students explained that in the interest of avoiding conflict or being stereotyped, they generally shunned all potentially contentious political discussion. By doing so, it should be apparent, the students also forfeited the opportunity to learn the true complexity (and the shared ambivalence) in each other's views. The obvious antidote to naive realism and its attributional consequences—that is, the open, sustained, sympathetic sharing of views and perspectives—was rarely employed by the students. Ironically, in attempting to avoid discomfort and giving offense, many students failed to discover that their particular position on the political spectrum (i.e., that of self-labeled realistic liberal or compassionate conservative) was one shared by a great number of their peers.

Apparent versus Real Differences in Basic Values and Golden Rules

Participants in political debates about issues that evoke strong feelings are bound to find others speaking and acting in ways that seem to reflect relative indifference to values that they themselves hold to be of paramount importance. As our discussion of false polarization suggests, such assessments are apt to be unwarranted; that is, the opposing participants infer (or tacitly assume without much deeper thought) that the other side's position reflects rejection of the values and standards that have dictated their own position. Thus advocates and opponents of universal health care, or of restrictive abortion laws, or of capital punishment, or of particular affirmative action policies, are apt to assume, often wrongly, that their adversaries lack compassion or reverence for life, or commitment to fairness, or insistence on personal responsibility, or some other generally shared value. What they fail to recognize is the extent to which their ideological opponents proceed not from radically different values and priorities but from differing factual assumptions and construals, and more importantly, from very different perceptions concerning the links between relevant perceptions, political positions, and values.

Thus death-penalty opponents and proponents differ in what they assume to be true about the causes of murder, the equity of sentencing, and the deterrent value of executions. And opponents and proponents of more restrictive abortion laws may differ in what they assume about the life circumstance and moral calculations of most women who choose abortion. Regardless of whether such differing assumptions and construals are the source of the relevant policy disagreements or rationalizations for positions that actually reflect other factors, mutual misattributions are likely to ensue. Each side sees their own position as the reasonable and ethical one and the opposite

position as *morally* and *ethically*, as well as *pragmatically*, deficient.

As I noted earlier, those who hold the conviction that the other side has acted out of pure self-interest or ideological bias, and has done so with little concern for or appreciation of universal values, often cannot help but notice that *other* people's views, construals, assumptions, and political positions generally do prove to be suspiciously congruent with their overall ideology and their personal or group interests. Furthermore, those on the other side (and even those on one's own side) seem disinclined to express the kinds of reservations and sources of ambivalence that accompany one's own views. What people often fail to note, of course, is that the same congruence can be found with respect to their *own* assumptions, interests, and beliefs. Moreover, they fail to consider the extent to which other individuals—both on their own side and on the other side of the debate—may, like them, hold more ambivalent and complex views than they are comfortable about expressing to anyone but trusted intimates. Thus, when the naive realist hears spokespersons for the other side support their position with appeals to universal values such as equality, justice, self-determination, reverence for life, or compassion for those victimized by the status quo, the appeals are seen as cynical or at best misguided. The *real* situation according to the naive realist—at least when it is appraised dispassionately (i.e., as the naive realist appraises it)—could lead the ethical actor and possessor of universal values to only one position, that is, the position that the naive realist happens to hold. In a sense, the failure is one of attributional "charity."

Social perceivers, as we have noted, have come to recognize that different actors not only have different preferences or tastes but also different perspectives and perceptions, and as such, that their own construals or constructions of social actions and entities may not be shared by their peers. Such insights about the diversity of subjective responses can, of course, be very helpful in promoting more accurate social predictions and inferences. Nevertheless, as we suggested earlier, the wise social perceiver should, at least tentatively, treat surprising, seemingly inappropriate or counternormative responses on the part of others as *symptoms* of such construal differences rather than uncharitably inferring negative personal traits (or, we would now add, inferring deficiencies or dramatic differences in personal values). In short, the naive conviction that others share our way of responding to the world—when adopted mindfully and selectively rather than assumed mindlessly and indiscriminately—can often be helpful in sparing us premature and erroneous assumptions about the values adhered to by others.

The English philosopher Thomas Hobbes offered the following prescription assumptions about others' subjective responses:

> Given the similitude of the thoughts and passions of one man to the thoughts and passions of another, whosoever looketh into himself and considereth what he doth when he does think, opine, reason, hope, fear, etc., and upon what grounds, he shall thereby read and know what are the thoughts and passions of all other men upon the like occasions (cited by Leakey and Lewin, 1992).

While the assumption that others share our basic values and preferences can shield us from various errors of inference and prediction, it can also prompt unwelcome paternalism and proselytism. The so-called Golden Rule, which is an essential feature not only of Christianity but of virtually all of the world's major religions, holds that we should do unto others as we would have them do unto us. But the English playwright George Bernard Shaw offers the following maxim: *Do not do unto others as you have them do unto you; their tastes may be different.* And even if their tastes do not differ from your own, their standards and priorities may differ markedly. A less presumptuous version of the Golden Rule espoused in various religions (and one that Shaw might even have endorsed) holds that we should *not* do unto others what we would *not* have others do unto us.

The comedian George Carlin offered a similarly acute psychological observation. Carlin asked the members of his audience, "Ever notice that anyone going slower than you is an idiot and anyone going faster is a maniac?" When first confronted with the observation, which pertained to freeway driving, we are inclined to chuckle, but we recognize that most of the time we do feel that those driving faster than us are reckless and those driving slower are overly cautious or perhaps just distracted. The same negative attributions occur as people respond to each others' words and deeds in the political domain that reflect views about which problems merit attention and the commitment of resources, and especially how fast we should move and how much we should spend in addressing those problems.

Nevertheless a reasonable prescription would be to assume, at least tentatively, that others, like you, value friendship and family highly; that they, like you, believe that justice should be served (albeit tempered with mercy); and that the other values you regard as essential to fair and ethical conduct are shared (although perhaps not ordered identically) by your peers and adversaries alike. Moreover, even when others

respond in a way that seems unreasonable, unconscionable, or simply bizarre, you should not give up such charitable assumptions unless and until you have ruled out the possibility that those others have proceeded not from values that differ markedly from your own, but rather from very different construals or interpretations of the relevant objects of evaluation and their relevance to such values.

Barriers to Dispute Resolution and Efficient Agreements

Negotiation processes and outcomes are the product of many decisions, in particular, whether to invite the other side to the negotiation table (or accept their invitation to meet) and what to say and do before, during, and after the negotiation. To some extent, these decisions reflect rational calculations of present and future interests, calculations that are made in light of existing relationships and prior history between the parties. But such calculations, and sometimes the failure to engage in them, are mediated by cognitive, motivational, and social processes that are the focus of contemporary theory and research in many social science disciplines, including psychology, political science, and economics.

Sometimes, of course, the interests of the parties are so diametrically opposed that the minimal needs and demands of the two parties cannot be reconciled, and a stalemate is inevitable until the situation changes, or until one side is able to impose its will on the other. But sometimes parties fail to negotiate, or negotiations end in failure, even when third parties and, indeed, even when representatives or majorities within the principle parties themselves can see shared interests and envision agreements that would leave all concerned better off than maintenance of the status quo. To understand such failures, to understand why the parties continue to bear the costs and uncertainties of ongoing struggle, we must look beyond an analysis of objective interests and beyond barriers to agreement arising from tactical or strategies blunders, or political constraints, or the self-interests of agents or elites or factions that can exercise veto power. (For a discussion of strategic, structural, and other nonpsychological biases, see Ross and Stillinger, 1991; also Mnookin and Ross, 1995.) We must examine the ways parties think and feel about their conflict and about each other and about the way they assign responsibility or blame for the conflict and past failures to resolve it, including perceptions and assessments in the course of the negotiation process itself.

Our preceding discussion of biases in construal and of naive realism provides a starting point for such an examination. The features of egocentrism and naive realism described so far not only give rise to misattribution, mistrust, and unwarranted pessimism about the prospects for finding common ground, they also create barriers to successful negotiation and dispute resolution. One such barrier—the false polarization phenomenon, and the concomitant underestimation of common ground, which makes parties skeptical about the possibilities of finding reasonable, pragmatic, counterparts on the other side of the conflict—has already been noted. But several other psychological barriers can thwart the efforts of negotiators and would-be peacemakers, some of which shall now be considered in some detail.

Cognitive Dissonance and Rationalization

One of the best-known and most extensively researched psychological biases involves motivated effort to seek and preserve cognitive consistency, and conversely to avoid and reduce *dissonance*, vis-a-vis one's actions, values, feelings, or beliefs (Festinger, 1957). Upon reflection, however, it becomes clear that sometimes the beliefs and priorities undertaken to reduce dissonance can create a stumbling block to dispute resolution. That is, past justifications and rationalizations offered to others and to the self for continuing the struggle (the other side is the devil incarnate; we can't trust him; God/ history is on our side; we are more resolute than the other side because right makes might; we can't break faith with the martyrs who gave their lives; etc.) increase the psychic and social costs of giving up the struggle and accepting the terms now on the table. This process is especially apt to make its influence felt when the conflictual and status quo has been costly to both sides and maintained for long periods of time and when possible resolutions offering apparent advantages over the status quo have, for one reason or another, been rejected in the past.

While the implications of dissonance reduction may be bleak in the context of protracted and costly stalemates, there is one optimistic note worth sounding before I continue my account of the formidable barriers that stand in the way of bargaining efficiency and success. Once a settlement *has* been reached, the process of dissonance reduction can play a somewhat constructive role—especially if the decision to settle has been freely reached, if effort has been expended or sacrifices made, and if public defense of the settlement has been demanded (Aronson, 1969; Brehm and Cohen, 1962). Thus, leaders and followers alike

may strive to find and exaggerate positive features and unanticipated benefits of the settlement and to minimize or disregard negative ones. We saw such processes occur in dramatic fashion early in 1972 when Richard Nixon suddenly and unexpectedly reached detente with China. And we have some optimism that the same processes will operate in the aftermath of agreements in the Middle East and other troubled areas of the globe.

The Pursuit of Fairness, Equity, or Justice

Negotiating parties seek to engage in exchanges whereby each improves their situation by ceding things they value less than the other party in order to gain things they value more than the other party. But in the context of long-standing conflict, the parties typically seek more than a simple advance over the status quo—they demand, and feel entitled, to receive *fairness*, or even true *justice* (see Adams, 1965; Homans, 1961; also Walster, Berscheid, and Walster, 1973). The parties want an agreement that allocates gains and losses in a manner proportionate to the strength and legitimacy of their respective claims (see Bazerman, White, and Loewenstein, 1995). Such demands raise the bar for the negotiators, especially when the parties inevitably have different narratives about past events and thus differ as to what an equitable agreement would be. Indeed, those differences in narratives and demands, when viewed by the parties through the prism of naive realism, engender further pessimism and distrust.

Both sides in the conflict feel that it is *they* who have acted more honorably in the past, *they* who have been more sinned against than sinning, and *they* who are seeking no more than that to which they are entitled. They forget that in their own families and communities, they continually tolerate outcomes that they deem unfair in the interests of harmony and positive relationships. Both sides, moreover, are apt to feel that it is *their* interests that most require protection in any negotiated agreement—for example, by avoiding ambiguities in language that could provide loopholes that could be exploited by the other side (while, at the same time, avoiding unrealistically rigid requirements for their own side that could compromise their ability to deal with unforeseen future developments). They also are bound to have divergent views about the future (i.e., who will grow stronger with the passage of time and whose assurances can be taken at face value and trusted).

Third-party mediators may face a particularly difficult challenge. The "impartial" civilian review board proposed by the mayor's task force to deal with allegations of racist-inspired police brutality is apt to be distrusted both by outraged members of the minority community (who fear it will be composed of middle-class whites who have never faced ill-treatment from the police and thus will take the word of a white police over "folks like us") and by the skeptical and beleaguered police officers (who fear it will be composed of civilians who do not understand our problems and frustrations and of political hacks who will try to placate voters). The willingness to accept such a review board, accordingly, would be seen by each side as a major concession to the other side. And when each side hears the other side's characterization of the content and equitability of the proposal, the result is likely to be heightened enmity and distrust.

Similar processes make the parties disagree strongly about the balance of any proposal that seeks to give both parties what they feel they need and deserve. Moreover, the disputants are apt to misattribute each others' cool response to the proposal in such a way that heightens enmity and mistrust. Each party is likely to feel that the other is being disingenuous in its public pronouncement of concern and disappointment and that the other is merely engaging in "strategic" behavior designed to secure sympathy from third parties and win further concessions. And, of course, each party responds with anger and suspicion when it hears its own response characterized in such uncharitable terms.

No laboratory experiment is required to demonstrate the pattern of costly stalemates, misattributions, and ever-growing enmity predicted by our analysis. The news media, with their continual accounts of ethnic strife and intergroup conflict, provide all the evidence one could wish. However, a pair of laboratory studies (Diekmann et al., 1997) conducted with students at Northwestern and Stanford Universities offers a subtler, and more hopeful, account of the interplay between equity concerns and self-interested construal biases. The results of these studies, which involved negotiations about the (hypothetical) allocation of resources, suggest that biases reflecting self-interest can be trumped by fairness norms, especially norms that promote equal treatment—but only when neither party has already enjoyed or been "endowed" with superior rights or treatment. When the starting point is one of inequality, those already advantaged prove quite willing and able to justify the relevant inequality of treatment—an inequality that few personally would have recommended, demanded, or imposed if they had not already been granted favored treatment.

To some extent, the findings by Diekmann et al. anticipate a phenomenon, and potential impediment to negotiated agreements, to be discussed next—that is, the phenomenon of *loss aversion*, which has

been documented by Kahneman and Tversky (1979, 1984) in their accounts of *prospect theory*. According to that theory, the prospect of a given loss is deemed more unattractive than the prospect of a gain of the same magnitude is deemed attractive, and as a result, people show themselves more motivated to avoid the loss than they are motivated to achieve the objectively equivalent gain. There is also a footnote to be added to this account of the Diekmann et al. study. When asked to predict how *others* in the study would respond to the same allocation decision, the research participants proved to be overly cynical and uncharitable in their predictions. They greatly overestimated the degree of partisan bias that other allocators and evaluators would show. The leap from this result in the context of a hypothetical negotiation exercise to predictions about more difficult, real-world negotiations is a large one. But the possibility that parties negotiating from a position of equality (in the absence of "facts created on the ground" that benefit one side or other) may actually *over*estimate the difficulty of reaching a mutually acceptable agreement offers one bright spot in an otherwise largely gloomy picture.

Reactive Devaluation, Loss Aversion, and Reluctance to Trade

Beyond the impediments to negotiated agreement posed by the motivational and cognitive biases discussed thus far, there is a further barrier resulting from the dynamics of the negotiation process itself that has been documented in research. That is, the evaluation of specific package deals and compromises may *change* as a consequence of their having been put on the table, especially if they have been offered or proposed by one's adversary. Evidence for such reactive devaluation has been provided in laboratory and field settings in which subjects evaluated a variety of actual or hypothetical dispute-resolution contexts and proposals (Ross and Ward, 1995).

Three findings emerge from this work. First, and perhaps least surprising, the terms of a compromise proposal for bilateral concessions are rated less positively when they have been put forward by the other side than when the same terms ostensibly have been put forward by a representative of one's own side. This was demonstrated convincingly in a study by Maoz et al. (2002) in which Israeli Arabs and Jews rated actual proposals put forward by the two sides in the post-Oslo negotiations, with the putative authorship of those proposals manipulated by the experimenter. As predicted, the putative authorship influenced the relative attractiveness of these proposals to the two groups of participants. Indeed, when the Israeli proposal was attributed to the Palestinian side in the negotiation and vice versa, the Israeli participants rated the actual proposal of their side to be less attractive than the actual proposal of the other side.

Two less obvious findings demonstrating reactive devaluation were reported in a study (Ross, 1995; see also Ross and Ward, 1995) in which the research participants were Stanford University students who wanted the university's immediate and total divestment of all shares held in companies doing business in the then-apartheid regime of South Africa. During the course of ongoing negotiations, students were asked to evaluate less radical proposals—first, those under discussion (i.e., partial or selective divestment from companies directly supporting the regime and its policies versus total divestment at a later date if particular policy changes were not made) and later, the plan (a version of the partial divestment) ultimately adopted by the university. As predicted, the participants rated more favorably whichever of two proposals that they were led to believe was about to be put forward than they rated the alternative. And when the university finally acted, its proposal was rated less positively than it had been when it was merely one of two possibilities. Conversely, a previously unpopular alternative calling merely for increased investment in companies that had elected not to do business in South Africa was rated more positively than it had been before the university's decision.

The latter findings may reflect a more general tendency for people to devalue that which is at hand or readily available relative to that which seems unavailable or is withheld (Brehm, 1966; Wicklund, 1974) because of a desire to maintain future freedom of action. But the phenomenon of *loss aversion* also may play an important role; that is, to the extent that a negotiation proposal represents a proposed trade of concession, the prospective losses that will be incurred by the acceptance of that proposal may loom larger and receive more weight than the prospective gains. Proposals that involve acceptance of negative change and/or heightened risk in return for prospective, or even certain, gains are particularly likely to be treated with caution. The literature on community members' willingness to pay to reduce environmental nuisances and risk versus their willingness to accept payment to accept such risk or nuisance, and the so-called NIMBY (not-in-my-backyard) phenomenon is instructive. A typical result is that communities demand payments many times greater to accept a new toxic-waste site, halfway house, or other unattractive change to their communities than they would be willing to pay to eliminate the existing unattractive presences.

A simple thought experiment, however, can illustrate the phenomenon to the reader. Consider the answers to the following three questions. How

much would you be willing to pay someone (perhaps a manufacturer or researcher on auto safety) to make the brakes in your automobile 10% *safer*? How large a payment would you demand to let someone (perhaps the same manufacturer or researcher) make your brakes 10% *less safe*? Finally, exactly how safe are your brakes right now compared with those in other cars or compared with how safe they could be made by a skilled mechanic or through installation of some new features? If you are a typical motorist, the answers to the first two questions will be (1) a reasonable but not a huge amount—somewhere between $100 and $1,000 and (2) a huge amount—at least many thousands of dollars—or, in many cases, no amount will be enough. For most people, ironically, the answer to the third question will be—"I guess they are pretty safe, but I don't really know how they compare to other folks' brakes or how much safer they could be made if I were willing to spend some money."

Other mechanisms may contribute to reactive devaluation as well. Concession offers may lead the recipients to conclude that the other side is eager for an agreement; that one's previous negotiation stance has been too moderate and that more can be extracted from the other side; and that one can win those concessions at a smaller price. Social processes can be involved as well. Critics who oppose agreement because they prefer conflict or have something to lose from agreement will inevitably dismiss preliminary proposals, or even unilateral concessions, as trivial, token, and insincere. But regardless of why reactive devaluation occurs, its potential contribution to the maintenance of negotiation deadlocks and to the ensuing cycle of heightening enmity and mistrust should be clear. Not only are proposals likely to be received less positively than they ought to be in terms of the objective interests of the parties, but each side is apt to interpret the other side's negotiation behavior and rhetoric as at best strategic manipulation, and at worst as dishonest, cynical, and dictated by animus rather than a sincere effort to end the conflict.

In the author's experience, seasoned negotiators are well aware of the reactive devaluation phenomenon, although they may not recognize its ubiquity or understand all of the psychological mechanisms that underlie it (or perhaps most importantly, recognize their own susceptibility to it). Indeed, one important role played by the mediator in any conflict is to short-circuit this process—to obscure the parentage of specific proposals and concessions and to encourage more positive (and accurate) attributions on the part of the disputants as they struggle to reach terms of agreement that are personally and politically bearable. To this end, skilled mediators may oblige the disputants to clarify their priorities and interests; in particular, to have each side indicate the concessions it may value more highly than its adversaries' and vice versa. The mediator then is free to propose possible exchanges of concessions that are not only based on but also readily attributable by the parties to their own particular expressions of priority.

Implications for Intergroup Dialogue and Conflict Resolution

The study of barriers and biases and of the psychological processes that exacerbate them can help us understand why intergroup dialogue so often proves frustrating for the participants and why negotiations sometimes fail when an objective analysis of the mutual advantages of changes to status suggests they should succeed. Such study can also help us understand why the process of negotiation, even when undertaken with great sincerity and motivation to reach agreement, can sometimes escalate, rather than attenuate, feelings of enmity and mistrust. But it can do more. It can contribute to the analysis and development of techniques for reducing misunderstanding and facilitating successful negotiation.

A comprehensive discussion of this topic is not possible here, but in the remainder of this chapter, I will outline some specific methods that mediators, facilitators, and other third parties, and to some extent enlightened citizens within divided societies, can employ. My emphasis primarily will be on results from laboratory and field studies that suggest potential techniques to counteract the biases and barriers discussed in this chapter, but I will also include some lessons from my own real-world experience in working with my SCICN colleagues to promote fruitful citizen dialogue and facilitate efforts at second-track diplomacy in Northern Ireland and the Middle East.

Countering False Polarization and Discovering Common Ground

Third parties can make an important contribution both by helping to build trusting relationships and by shifting the focus of discussion from areas of disagreement to areas of common ground. This role is particularly important in mediating conflicts between groups or peoples with long histories of enmity. Indeed, in such cases there is obvious value merely in bringing disputants into sustained personal contact, so that they can get beyond stereotypes and discover shared fears and aspirations (Doob and Foltz, 1973; Kelman, 1995, 1997, 2001). The goal ultimately is to have the participants engage in frank, open dia-

logue—dialogue in which they talk about their factual assumptions and the complexities of their values rather than simply defending their positions.

Even if such dialogue does not lead to agreement, it is likely at least to challenge the partisans' view of the other side as monolithic, unreasoning, unreasonable, entirely ideologically driven, and unwilling to consider compromise. Such discussions can also help participants to see the inconsistencies, uncertainties, and disagreement in their own side's position, thus making them freer to express dissent and entertain new ideas. Indeed, it has been proposed that like third parties (Rubin, 1980), moderates within the rival groups (Jacobson, 1981) also have a valuable role to play in this regard in encouraging partisans to get beyond rhetoric and statements of position to the point of discussing underlying interests, assumptions, and concerns (see also Fisher, Ury, and Patton, 1991; Rubin, Pruitt, and Kim, 1994).

A particularly welcome development has been the flourishing of multitrack diplomacy (Diamond and McDonald, 1991) and public peace processes (Saunders, 1999), whereby peace-making or dispute-resolution initiatives are pursued by nongovernmental organizations, such as peace centers, academic institutions, and think-tanks; by representatives of particular professional organizations, such as physicians, social workers, or educators; and even by groups of concerned individuals. In fact, the path to the dramatic 1993 Israeli-Palestinian accord on Jericho and Gaza was blazed through such initiatives (Rouhana and Kelman, 1994). In our SCICN initiatives in the Middle East and Northern Ireland, we have seen participants build personal trust and rapport as the conversation gradually shifted from disputes about past wrongdoing to the personal histories that brought the participants to the table, and from charges and countercharges about the present conflict to the participants' vision of the type of society and shared future that they would like to create for their grandchildren.

Beyond seeing the development of warm personal relations, we also have noticed several less obvious benefits from these unofficial, nongovernmental initiatives. First, because the participants are not diplomats acting in an official capacity, they are free to explore new and visionary proposals without fearing the risks or spontaneity or the strategic disadvantages of candor and spontaneity. Second, because the meetings are informal, free from the glare of media, and relatively leisurely in their pace, the participants typically can socialize and talk about their families and personal experiences as they take walks and enjoy meals (and, in our conferences, even wash dishes) together, thereby creating personal bonds of respect and friendship that facilitate future contact and joint initiatives. Finally, and perhaps most importantly, the participants provide each other with invaluable information about priorities and areas of flexibility and potential movement—areas where exchanges of concessions could be sought. By exploring the views of moderates and peace advocates in the other camp, the participants gain a clearer sense of which of the other side's current demands reflect deeply held and widely shared sentiments and which reflect potential areas for negotiation and exchanges of concessions. These assessments, in turn, get more broadly disseminated when the citizen diplomats return home and share their experiences and impressions with influential colleagues and news media.

Framing and Construal

Framing can be an important tool in overcoming barriers to negotiation. When loss aversion and reactance are making parties respond negatively to proposed changes in the status quo, parties need to recognize that doing nothing is itself a decision, one that imposes costs and creates risks that are substantial and likely to increase in the future. Beyond focusing on the risks and cost of inaction, mediators can reduce the tendency for the bargainers to view their concessions as losses by inducing them to view the things they will be giving up as bargaining chips, or negotiation currency—something to be exchanged readily (in the way that one spends money or any other kind of currency) for things that one values more.

More important, perhaps, third parties can help the parties to overcome various psychological barriers in the way that they frame, or if necessary *re*frame the negotiation process itself. For example, in conflicts where both sides harbor long-standing, and often quite justified, grievances, they can frame the endeavor not as an attempt to redress injustice or to have the parties get what they are due in light of the strength and legitimacy of their claims, but rather as an exercise in problem solving. That is, the negotiator's task can be framed explicitly as one of discovering and exploiting opportunities for efficient exchange or mutually beneficial trades of concessions in light of the parties' differing needs and priorities (Raiffa, 1982; Rubin, Pruitt, and Kim, 1994). In any case, the framing of the negotiators' task should be one that encourages the parties to move beyond political posturing and recriminations about past wrongs to a concern with the future and the pursuit of enlightened self-interest or group interest.

While the leap from game-theory demonstrations in the laboratory to negotiation outcomes in the real world is a large one, I can report a study on the power of simple semantic framing on participants' responses

in the much-studied Prisoner's Dilemma game (Liberman, Samuels, and Ross, 2004). The framing manipulation in the study, which was conducted both at Stanford University and at an Israeli University that prepares students for careers in business and law, was a very simple one. On two occasions, in explaining the nature of the game and the relevant payoff matrix, the experimenter referred to the game either as the *Wall Street Game* or the *Community Game* (or, in the Israeli version, as the *Bursa* Game versus the *Kommuna* Game). The results of the manipulation were dramatic.

Only about one-third of players elected to cooperate in the first round of the *Wall Street* or *Bursa* Game, whereas more than two-thirds elected to cooperate in the first round of the *Community* or *Kommuna Game*. These differences, moreover, persisted on subsequent rounds. Interestingly, while the label attached to the game seemed to play a large role in the way participants played, and also in the way they expected their counterparts to play, the participants' personal reputation for past cooperativeness or competitiveness had little predictive validity. Participants nominated by their dormitory advisors (to whom the game had been described in detail) as most likely to cooperate versus those nominated as least likely to cooperate did not differ at all in their first-round or subsequent decisions. It is not unreasonable to suggest that negotiators, as well, may sometimes be induced, by straightforward instructions and/or subtler features of the negotiation context (Kay and Ross, 2003), to regard their task as a cooperative undertaking to seek mutual advantage rather than as a competitive exercise, with some assurance that their counterpart is doing likewise, and that doing so will enhance the prospects for success.

Managing Attributions

Understanding the role that attributions play in reactive devaluation may be a key to overcoming this barrier to agreement. The relevant devaluation in question stems, at least in part, from people's basic tendency to search for causal explanations for each other's behavior (Heider, 1958; Jones and Davis, 1965; Kelley, 1967, 1973). Parties receiving proposals from the other side, especially in the context of conflict where there exists a history of enmity and failed past negotiations, are bound to search for explanations for the content and timing of such proposals. The recipient of a unilateral concession or proposed trade of concessions wonders, "Why is my adversary offering this *particular* concession or proposing this *particular* trade, and why *now*?" In the absence of other satisfactory answers, the recipient is likely to provide ones that lessen the chances for agreement; for example, that the concession offered is probably less substantial than it might seem on the surface, that the concession sought is more substantial and valuable than the recipient has previously recognized, or worse still, that the other side does not intend to keep its side of the bargain and will renege on its commitments once it has gotten what it wants.

The third-party mediator can sometimes help solve these attribution problems (and, in a sense, also help to reduce some of the dissonance felt by the negotiating parties) by pointing out, or by having the party offering the proposal indicate, to the recipient the political realities and constraints motivating the offer. Recognition of attribution considerations also suggests a potentially beneficial role of overt, external resolution pressures, such as deadlines or promises of side payments by interested third parties, in the negotiation process. The recipient of an attractive package offer or pump-priming concession is thereby provided with an *explanation* for the adversary's newly demonstrated flexibility, one that precludes the need for a *negative* reassessment of the significance of the various concessions offered and the exchanges proposed. Moreover, the existence of deadlines and other external resolution pressures provides the negotiating parties with more palatable attributions for their *own* flexibility (good sense, not weakness or gullibility) and better explanations to offer both the constituencies they represent and the critics they must face.

While the emphasis here has been on the role that mediators, facilitators, and other third parties can play in dealing with attributional problems in negotiation contexts, the principal parties themselves can address such problems. In particular, they can make some effort to link the content of their proposals to the specific pressures and constraints they are facing or, better still, to the expressed needs and desires of the other side. A recent demonstration experiment by Ward et al. (2008) illustrated this possibility. In the study, students negotiated with an experimental confederate regarding the recommendations to be made by their university with regard to the reform of drug laws. At a late stage in the negotiation, the confederate offered a compromise proposal (one calling for the legalization of marijuana but a harsher penalty for harder drugs and a provision to adopt more draconian measures if drug use increased rather than decreased over the course of the trial period).

The key finding was that when the confederate introduced his proposal by saying "I have heard your arguments and proposal so I am discarding the proposal I came with and offering this new proposal instead," it was rated more positively and more likely to be accepted by the students than when he said "Here

is the proposal I have brought to the negotiation." In other words the same proposal was received more positively when it was explicitly acknowledged as a change in position, one linked to the previously expressed priorities of the party to whom it was offered. Such acknowledgment, it should be noted, is often absent in the context of ongoing conflicts, at least in public pronouncements wherein both of the parties in the conflict assure their constituencies that they are holding firm to their longstanding position.

Negotiation Expectations and Ideology

Consideration of conflict-resolution efforts that fail even though the cost is the prolonging of a debilitating conflict, or that succeed even when common ground seems difficult to find, would be incomplete without some acknowledgement of the role of expectation. In particular, experience suggests that agreement is likely to be achieved, and in fact can become a virtual certainty, when the opposing parties enter the negotiation process with a historically based expectation, and the absolute conviction that such resolution can, must, and will be achieved (and with confidence that their adversaries share such expectations and convictions). In contrast, the achievement of agreement seems to become unlikely when (and one could argue *because*) the parties have a history of prior failures, and, honorable intentions and ample motivation notwithstanding, they enter into the negotiation with grave doubts about the possibility of achieving any major breakthroughs in the question for agreement.

When the absolute necessity and the absolute *inevitability* of resolution is accepted by both parties, the possibility of success is greatly enhanced, not only because the parties have a powerful and shared motive to avoid failure, but also—as our analysis of the significance of attributional issues suggests—because they have a satisfactory explanation for any concessions that they are obliged to make and a satisfactory attribution for apparent concessions by the other side (i.e., "we gave in on some key points because we *had* to reach an agreement, and so did *they*"). Examples of difficult negotiations in which success is seen by the parties as inevitable in spite of the various barriers we have identified, and examples of such successes, which can be seen as self-fulfilling prophecies, are instructive (for some examples, see Ross and Stillinger, 1991). Some cases (such as the election of a pope) involve negotiations that occur irregularly and can be seen as exceptional events. Others (such as the passing of a federal or state budget) occur with great regularity and in a sense are seen as ordinary.

In both cases, the conflicts and divisions are complex and deep, and objective analysis of the interests of the contending factions might suggest that no solution could be proposed that would command a majority (much less the two-thirds majority required to elect a pope). But the certainty of relatively timely resolution, buttressed by history and tradition, seems to guarantee that a resolution will be found within the expected time period. That resolution will inevitably involve major compromises by, and the dissatisfaction of, those whose interests would seem better served by prolonging the negotiation in the hope of a better deal. But the agreement will be justified by the negotiators (to themselves, as well as to others) by a simple dictum: "we *had* to have a pope" or "we *had* to have a budget."

The implications of this discussion of negotiation ideology and expectation are worth underscoring. When obstacles are formidable and the sentiment exists that failure is possible or thinkable, such failure becomes highly likely. Only when failure is unthinkable can one be optimistic that the cognitive barriers discussed in this chapter will begin to crumble. There is no magic formula available to create such positive and self-fulfilling expectations, but the results of a pair of laboratory studies illustrate the potential value of engendering positive expectations on the part of negotiators (Liberman, Anderson, and Ross, 2009). One study was conducted in the United States with American college students doing a relatively mundane negotiation exercise involving the allocation of resources to undergraduate and graduate student activities. The other was conducted in Israel with young people who had served in the military negotiating a structurally similar, but much more politically sensitive, resource allocation problem that involved funds associated with the building of the fence or wall separating the West Bank Jews from the Palestinians.

In both cases, the experimental procedure involved the use of a confederate who followed a predetermined script. In the U.S. study, the confederate was an older student ostensibly representing the interests of the graduate students; in the Israeli study, it was an Arab confederate ostensibly representing the interests of the Palestinians. In both studies, the negotiation proceeded in stages, with the confederate making an initial offer, the experimental participant making a counteroffer, and the confederate making a final offer as time was expiring, The participants both assessed that offer and decided to accept or reject it knowing that the result of rejection would be a forfeiture of the funds in question until some later date. The experimental manipulation was a simple one. Half of the participants were informed at the outset of their negotiation that all previous negotiating pairs (in the case of the U.S. study) or "virtually all" (in the Israeli study) had succeeded in reaching agreement.

Half were given no such induction; they were merely told to do their best to reach an agreement. While the participants recognized the hypothetical nature of their role-play assignment, they did represent the interests of their own group and negotiated seriously, and in some cases, especially in the Israeli study, quite passionately.

As predicted, the manipulation had a dramatic effect on the participants. Prior information about the universal success of previous negotiators resulted in more generous counteroffers to the confederate's initial proposal, more positive assessments of the confederate's final offer, and unanimous acceptance (in the U.S. study) or near unanimous acceptance (in the Israeli study) of that proposal, whereas the same proposal was overwhelmingly rejected in both studies by the participants merely urged to do their best to reach an agreement. Moreover, participants in positive expectations conditions who reached agreement did not do so grudgingly. In fact, they subsequently provided more positive assessments of the negotiation process and of their counterpart than those who reached agreement under the control condition.

While this exact manipulation obviously would not be possible in most real-world situations, there is much third parties and the principal parties themselves can do and say to engender optimism about the outcome of sincere negotiation efforts. One possibility is to undertake a series of prior negotiations on issues that are relatively uncontroversial and yield obvious benefits to both sides. Another involves the value of words and deeds that defy expectations—as did President Sadat's dramatic 1977 trip to Jerusalem and speech before the Knesset in pursuit of peace between Israel and Egypt—in a way that suggests that things have changed and that past intransigency at the negotiation table will not be repeated.

Education: Benefits of Understanding the Underlying Processes

Consideration of practical measures to achieve conflict resolution prompts an optimistic hypothesis about the role of education and insight. Simply stated, the hypothesis is that awareness of the psychological processes outlined in this chapter—especially the processes of biased construal and reactive devaluation that make parties respond coolly to concessions, compromise packages, and other harmony-seeking overtures—will forestall them or weaken their impact.

In a sense, it is proposed that adversaries who have been "debriefed" about social and psychological processes that limit their capacity to negotiate compromises that are in their own best interests will have a useful tool for self-appraisal and for educating relevant constituencies (Ross, Lepper, and Hubbard, 1975). Such insights about process can also serve third-party mediators well—both in helping them to define their roles and in helping to overcome the pitfalls facing the adversaries. Finally, and most generally, education and insight about barriers to conflict resolution can aid people of goodwill in designing conflict-resolution strategies and tactics that allow the adversaries either to avoid the barriers in question or to overcome them.

When Experience Informs Theory and Applications: Four Lessons from the Real World

My own experience in dialogue and multi-track work and that of my colleagues at the SCICN has served mainly to reinforce fundamental, hard-won insights that social psychology has long offered to those who would listen. These lessons include the power of the situation in general and that of group norms and group pressures in particular; the need to attend to the actors' construal of their situation and the meaning they attach to their own actions and outcomes, as well as the meaning they attach to the behavior directed toward them by others; and the lengths to which people will go to in order to see themselves as rational and good (Ross, Lepper, and Ward, 2009). But at times, experience has alerted me to the importance of other, more specific influences. In each case I will simply state the lesson, and, where I think it might be helpful to the reader, say a bit about its origin and/or relevance.

The importance of a shared view of (and shared commitment to) a mutually bearable future. . . . Without it, negotiation between leaders and their agents, and even second-track diplomacy, is doomed to produce failure and to heighten rather than ease distrust. Repeatedly, we have found that parties come to dialogue groups and pursuit of common ground in order to specify what they need and want, why they are entitled to it, and what they are prepared to offer in return. What they typically fail to offer is a view of the future that specifies the place of those on the other side that is demonstrably better than the status quo. We have learned to challenge participants to explain why the life they envision for those on the other side is better than the other side imagines or fears it would be in any such agreement. If they can't offer such a view of a shared future, we suggest, there is no point in meeting, because that meeting will merely confirm the other sides' misgivings and strengthen its resolve.

The importance of relationships and trust—especially in dealing with spoilers and the demands of internal politics. Deals that make the larger populations on both sides of a conflict better off will almost

always leave some individuals or interest groups worse off. In the most virulent conflicts, some of these individuals and interest groups will go beyond the normal realm of political discourse to make their objections felt and do their utmost, including resorting to violence and intimidation, to ensure that no agreement will be reached. Parties seeking a lasting agreement must also agree on how to deal with such "spoilers." Moreover, the two parties cannot treat the efforts of spoilers on the other side, and the lack of quick and effective action against them, as evidence of bad faith while at the same time insisting that the other side should not be deterred by spoilers on their side and should understand that the demands of internal politics prevent them from cracking down on such spoilers.

The futility of trying to convince people of something they cannot "afford" to understand. One of the underexplored implications of dissonance theory (and psychodynamic theories, more generally) involves the limits to the value of appeals to reason or even ethical principles. When the threat of loss or the cost of an agreement to the self is too great, people will find a reason, or at least a rationalization, for continued intransigence. The cost in question may involve the need to recognize that one's life has been spent in a fruitless endeavor or that sacrifices of blood and treasury have been in vain. But it may simply involve the unacceptability of the life and status that awaits one post-agreement. I vividly remember a Protestant militia leader who had come out of prison ready to renounce violence and willing to negotiate earnestly with the other side. Yet, somehow, no deal put on the table was ever good enough, no promise by the other side reliable enough, to get him to say, "Let's stop talking and close the deal!" Observing this charismatic but uneducated man one could not escape the thought that right now he was a respected leader with a place at the negotiating table, but that in the aftermath of the agreement, and with the emergence of a normal peaceful society, he would be lucky to get a job driving a brewery truck. The issue of a bearable future pertains to individuals—especially those with veto power—as well as to groups and whole societies.

Conversion from militant to peacemaker need not involve any "blinding light" conversion. Sometimes it is "51% versus 49%." A common refrain we hear from moderates on one side looking for counterparts on the other side is that their side cannot make a deal with some particular leader, or that what they are waiting for is a Mandela on the other side. Rather than offering a windy lecture on the sins of dispositionism and the fundamental attribution error, we point out that Mandela was able to make peace not because he made the compromises that no one else would make, but because he made himself the leader with whom White South Africans were willing to make the compromises that they said *they* would never make (largely because he offered them a future in which they would have an acceptable place).

We also tell them about a provocative set of remarks that David Ervine, a Northern Ireland Loyalist and ex-bomber, made in an address at Stanford University in response to an inevitable question asking him about the insight or personal transformation that had changed him from a militant bomber to a mainstream politician determined to achieve a peaceful solution to the conflict. He explained that in his case it was a matter of "51% vs. 49%"—that his change involved not a transformation of character but a kind of "tipping point" whereby the futility and costs of violence became marginally more obvious and the prospects for securing an acceptable agreement through normal politics became marginally brighter. He then added the striking comment that when he was only 51% certain about the decision to embrace bombing as a tactic, he was still 100% a bomber, and now that he is only 51% certain about the prospects for change through peaceful means, he is 100% a politician and peace activist. The moral of this story is clear. Not only does the situation matter, but small changes for the better are worth working for. A meeting with the other side that goes well, a small concession that makes life for the other side more bearable, or a single humanizing remark can provide the tipping point that makes the difference between peace and conflict.

Note

This material is based upon work supported by the National Science Foundation under Grant No. 0447110. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Adams, J. S. (1965). Inequality in social exchange. In L. Berkowitz (Ed.), *Advances in social psychology* (Vol. 2, pp. 267–299). New York: Academic Press.

Aronson, E. (1969). A theory of cognitive dissonance. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 4, pp. 1–34). New York: Academic Press.

Asch, S. E. (1952). *Social psychology*. New York: Prentice-Hall.

Bazerman, M. H., White, S. B., and Loewenstein, G. F. (1995). Perceptions of fairness in interpersonal and individual choice situations. *Current Directions in Psychological Science*, *4*, 39–43.

Brehm, J. W. (1966). *A theory of psychological reactance*. Oxford, UK: Academic Press.

Brehm, J. W., and Cohen, A. R. (1962). *Explorations in cognitive dissonance*. New York: John Wiley and Sons.

Bruner, J. S. (1957). Going beyond the information given. In H. Gruber, K. R. Hammond, and R. Jesser (Eds.), *Contemporary approaches to cognition* (pp. 41–69). Cambridge, MA: Harvard University Press.

Bryan, C. J., Dweck, C. S., Ross, L., Kay, A. C, and Mislavsky, N. (2009). Ideology as mindset: Effects of personally-relevant priming on political assessments. *Journal of Experimental Social Psychology*, *45*, 890–895.

Cohen, S. P., Kelman, H. C., Miller, F. D., and Smith, B. L. (1977). Evolving intergroup techniques for conflict resolution: An Israeli-Palestinian pilot workshop. *Journal of Social Issues*, *33*(1), 165–189.

Davison, W. P. (1983). The third-person effect in communication. *Public Opinion Quarterly*, *47*(1), 1–15.

Diamond, L., and McDonald, J. (1991). *Multi-track diplomacy: A systems guide and analysis*. Iowa Peace Institute Occasional Paper 3. Iowa Peace Institute, Iowa.

Diekmann, K., Samuels, S., Ross, L., and Bazerman, M. (1997). Self-interest and fairness of resource allocation: Allocators versus recipients. *Journal of Personality and Social Psychology*, *72*, 1061–1074.

Doob, L., and Foltz, W. J. (1973). The Belfast Workshop: An application of group techniques to a destructive conflict. *Journal of Conflict Resolution*, *17*(3), 489–512.

Ehrlinger, J., Gilovich, T., and Ross, L. (2005). Peering into the bias blindspot: People's assessments of bias in themselves and others. *Personality and Social Psychology Bulletin*, *31*, 680–692.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Fisher, R. Ury, W., and Patton, B. (1991) *Getting to yes: Negotiating agreement without giving in* (2d ed). Boston: Houghton Mifflin Harcourt.

Fiske, S. T., and Taylor, S. E. (1984). *Social cognition*. New York: Random House.

———. (2008). *Social cognition: From brains to culture*. Boston, MA: McGraw-Hill.

Gilovich, T. (1981). Seeing the past in the present: The effect of associations to familiar events on judgments and decisions. *Journal of Personality and Social Psychology*, *40*, 797–808.

———. (1990). Differential construal and the false consensus effect. *Journal of Personality and Social Psychology*, *59*(4), 623–634.

Griffin, D.W., and Ross, L. (1991). Subjective construal, social inference, and human misunderstanding. In M. P. Zanna (Ed.). *Advances in experimental social psychology* (Vol. 21, pp. 319–359). New York: Academic Press.

Hastorf, A., and Cantril, H. (1954). They saw a game: A case study. *Journal of Abnormal and Social Psychology*, *49*, 129–134.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

Higgins, E. T., Rholes, W. S., and Jones, C. R. (1977). Category accessibility and impression formation. *Journal of Experimental Social Psychology*, *13*, 141–154.

Homans, G. (1961). *Social behaviour: Its elementary forms*. London: Routledge and Kegan Paul.

Ichheiser, G. (1949). Misunderstandings in human relations: A study in false social perception. *American Journal of Sociology* (Supplement), *55*, 1–70.

———. (1970). *Appearances and realities*. San Francisco: Jossey-Bass.

Jacobson, D. (1981). Intraparty dissensus and interparty conflict resolution: A laboratory experiment in the context of the middle east conflict. *Journal of Conflict Resolution*, *25*, 471–494.

Janis, I. L. (1972). *Victims of groupthink*. Boston, MA: Houghton Mifflin.

Janis, I. L., and Mann, L. (1977). *Decision making: A psychological analysis of conflict, choice, and commitment*. New York: Free Press.

Jones, E. E., and Davis, K. E. (1965). From acts to dispositions: The attribution process in social psychology. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 2, pp. 219–266). New York: Academic Press.

Kahneman, D., and Tversky, A. (1979). Prospect theory: An analysis of decisions under risk. *Econometrica*, *47*, 313–327.

———. (1984). Choices, values, and frames. *American Psychologist*, *39*, 341–350.

Kay, A. C., and L. Ross (2003). The perceptual push: The interplay of implicit cues and explicit situational construals in the Prisoner's Dilemma. *Journal of Experimental Social Psychology*, *39*, 634–643.

Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska symposium on motivation* (Vol. 15, pp. 192–241). Lincoln, NE: University of Nebraska.

———. (1973). The process of causal attribution. *American Psychologist*, *28*, 107–128.

Kelman, H. C. (1995). Contributions of an unofficial conflict resolution effort to the Israeli-Palestinian breakthrough. *Negotiation Journal*, *11*, 19–27.

———. (1997). Group processes in the resolution of international conflicts: Experiences from the Israeli-Palestinian case. *American Psychologist*, *52*, 212–220.

———. (2001). The role of national identity in conflict resolution: Experiences from Israeli-Palestinian problem-solving workshops. In R. D. Ashmore, L. Jussim, and D. Wilder (Eds.), *Social identity, intergroup conflict, and conflict reduction* (pp. 187–212). Oxford: Oxford University Press.

Leakey, R., and Lewin, R. (1992). *Origins reconsidered*. London: Little, Brown and Co.

Liberman, V., Anderson, N., and Ross, L.(2009). *Achieving*

difficult agreements: Effects of positive versus neutral expectations on negotiation processes and outcomes. Manuscript submitted for publication.

Liberman, V., Samuels, S., and Ross, L. (2004). The name of the game: Predictive of reputations vs. situational labels in determining Prisoner's Dilemma game moves power. *Personality and Social Psychology Bulletin*, *30*, 1175–1185.

Lord, C. G., Ross, L., and 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.

Maoz, I., Ward, A., Katz, M., and Ross, L. (2002). Reactive devaluation of an "Israeli" vs. a "Palestinian" peace proposal. *Journal of Conflict Resolution*, *46*, 515–546.

Marks, G., and Miller, N. (1987). Ten years of research on the false-consensus effect: An empirical and theoretical review. *Psychological Bulletin*, *102*(1), 72–90.

Mnookin, L., and Ross, L. (1995) Strategic, psychological, and institutional barriers: An introduction. In K. Arrow, R. Mnookin, L. Ross, A. Tversky, and R. Wilson (Eds.), *Barriers to conflict resolution*. New York: Norton.

Nisbett, R. E., and Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall.

Paul, B., Salwen, M. B., and Dupagne, M. (2000). The third-person effect: A meta-analysis of the perceptual hypothesis. *Mass Communication and Society*, *3*(1), 57–85.

Pronin, E., Gilovich, T., and Ross, L. (2004). Objectivity in the eye of the beholder: Divergent perceptions of bias in self versus others. *Psychological Review*, *111*, 781–799.

Raiffa, H. (1982). *The art and science of negotiation*. Cambridge, MA: Harvard University Press.

Robinson, R., Keltner, D., Ward, A., and Ross, L. (1995). Actual versus assumed differences in construal: "Naive realism" in intergroup perception and conflict. *Journal of Personality and Social Psychology*, *68*, 404–417.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 10, pp. 173–240). New York: Academic Press.

———. (1995). The reactive devaluation barrier to dispute resolution. In K. Arrow, R. Mnookin, L. Ross, A. Tversky, and R. Wilson (Eds.), *Barriers to conflict resolution*. New York: Norton.

Ross, L., Greene, D., and House, P. (1977). The false consensus effect: An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, *13*, 279–301.

Ross, L., Lepper, M. R., and Hubbard, M. (1975). Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm. *Journal of Personality and Social Psychology*, *32*, 880–892.

Ross, L., Lepper, M. R., and Ward, A. (2009). A history of social psychology: Insights, contributions, and challenges. In S. Fiske and D. Gilbert (Eds.), *Handbook of social psychology* (4th ed., Vol. 1). New York: Random House.

Ross, L., and Nisbett, R. E. (1991). *The person and the situation: Perspectives of social psychology*. New York: McGraw-Hill.

Ross, L., and Stillinger, C. (1991). Barriers to conflict resolution. *Negotiation Journal*, *7*, 389–404.

Ross, L., and Ward, A. (1995). Psychological barriers to dispute resolution. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 27, pp. 255–304). San Diego, CA: Academic Press.

———. (1996). Naive realism in everyday life: Implications for social conflict and misunderstanding. In E. S. Reed, E. Turiel, and T. Brown (Eds.), *Values and knowledge* (pp. 103–135). Hillsdale, NJ: Erlbaum.

Rouhana, N. N., and Kelman, H. C. (1994). Promoting joint thinking in international conflicts: An Israeli-Palestinian continuing workshop. *Journal of Social Issues*, *50*, 157–178.

Rubin, J. Z. (1980). Experimental research on third-party intervention in conflict: Toward some generalizations. *Psychological Bulletin*, *87*(2), 379–391. doi:10.1037/0033-2909.87.2.379

Rubin, J. Z., Pruitt, D. G., and Kim, S. H. (1994). *Social conflict: Escalation, stalemate, and settlement* (2nd ed). New York: McGraw-Hill.

Saunders, H. S. (1999). *A public peace process: Sustained dialogue to transform racial and ethnic conflicts*. New York: St. Martin's Press.

Vallone, R. P., Ross, L., and Lepper, M. R. (1985). The hostile media phenomenon: Biased perception and perceptions of media bias in coverage of the Beirut massacre. *Journal of Personality and Social Psychology*, *49*, 577–585.

Walster, W. E., Bersheid, E., and Walster, G. W. (1973). New directions in equity research. *Journal of Personality and Social Psychology*, *25*, 151–176.

Ward, A., Disston, L. G., Brenner, L., and Ross, L. (2008). Acknowledging the other side in negotiation. *Negotiation Journal*, *24*, 269–285.

Wicklund, R. (1974). *Freedom and reactance*. Potomac, MD: John Wiley and Sons.

CHAPTER 7

Psychic Numbing and Mass Atrocity

PAUL SLOVIC

DAVID ZIONTS

ANDREW K. WOODS

RYAN GOODMAN

DEREK JINKS

> The means for expressing cruelty and carrying out mass killing have been fully developed. It is too late to stop the technology. It is to the psychology that we should now turn.
>
> —*Jonathan Glover, Humanity, 2001, p. 144*

The twentieth century is often said to have been the bloodiest century in recorded history. In addition to its wars, it witnessed many grave and widespread human rights abuses. But what stands out in historical accounts of those abuses, perhaps even more than the cruelty of their perpetration, is the inaction of bystanders. Why do people and their governments repeatedly fail to react to genocide and other mass-scale human rights violations?

There is no simple answer to this question. It is not because people are insensitive to the suffering of their fellow human beings—witness the extraordinary efforts an individual will expend to rescue a person in distress. It is not because people only care about identifiable victims of similar skin color who live nearby: witness the outpouring of aid from the north to the victims of the December 2004 tsunami in Southeast Asia. Nor can the blame be apportioned entirely to political leaders. Although President George W. Bush was unresponsive to the murder of hundreds of thousands of people in Darfur, it was his predecessor, President Bill Clinton, who ignored the genocide in Rwanda, and President Franklin D. Roosevelt who for too long did little to stop the Holocaust. The American example of inaction has been largely repeated in other countries as well. Behind every leader who ignored mass murder were millions of citizens whose indifference allowed the inaction to pass.

Every episode of mass murder is distinct and raises unique social, economic, military, and political obstacles to intervention. We therefore recognize that geopolitics, domestic politics, or failures of individual leadership have been important factors in particular episodes. But the repetitiveness of such atrocities, which have been ignored by powerful people and nations and by the general public, calls for explanations that may reflect some fundamental deficiency in our humanity—a deficiency not in our intentions, but in our very hardware, and a deficiency that once identified might possibly be overcome.

One fundamental mechanism that may play a role in many, if not all, episodes of mass-abuse neglect involves the capacity to experience *affect*, the positive and negative feelings that combine with reasoned analysis to guide our judgments, decisions, and actions. Research shows that the statistics of mass rights violations or genocide, no matter how large the numbers, fail to convey the true meaning of such atrocities. The numbers fail to spark emotion or feeling and thus fail to motivate action. The genocide in Darfur is real, but we do not "feel" that reality. We examine below ways that we might make genocide "feel real" and motivate appropriate interventions. Ultimately, however, we conclude that we cannot only depend on our intuitive feelings about these atrocities. In addition, we must create and commit ourselves to in-

stitutional, legal, and political responses based upon reasoned analysis of our moral obligations to stop large-scale human rights violations.

Lessons from Psychology

In 1994, Roméo Dallaire, the commander of the tiny UN peacekeeping mission in Rwanda, was forced to watch helplessly as the slaughter he had foreseen and warned about began to unfold. Writing of this massive humanitarian disaster a decade later, he encouraged scholars "to study this human tragedy and to contribute to our growing understanding of the genocide. If we do not understand what happened, how will we ever ensure it does not happen again?" (Dallaire, 2005, p. 548).

Researchers in psychology, economics, and a multidisciplinary field called behavioral decision theory have developed theories and findings that, in part, begin to explain the pervasive underresponse to atrocity.

Affect, Attention, Information, and Meaning

The search to identify a fundamental mechanism in human psychology that causes us to ignore mass murder and genocide draws upon a theoretical framework that describes the importance of emotions and feelings in guiding decision making and behavior. Perhaps the most basic form of feeling is affect, the sense (not necessarily conscious) that something is good or bad. Positive and negative feelings occur rapidly and automatically—note how quickly it takes to sense the feelings associated with the word *joy* or the word *hate*. A large research literature in psychology has documented the importance of affect in conveying meaning upon information and motivating behavior (Barrett and Salovey, 2002; Clark and Fiske, 1982; Forgas, 2000; Le Doux, 1996; Mowrer, 1960; Tomkins, 1962, 1963; Zajonc, 1980). Without affect, information lacks meaning and will not be used in judgment and decision making (Loewenstein et al., 2001; Slovic et al., 2002).

Affect plays a central role in what are known as "dual-process theories" of thinking. As Epstein (1994) observed, "There is no dearth of evidence in every day life that people apprehend reality in two fundamentally different ways, one variously labeled intuitive, automatic, natural, nonverbal, narrative, and experiential, and the other analytical, deliberative, verbal, and rational" (p. 710).

Stanovich and West (2000) labeled these two modes of thinking *System 1* and *System 2*. One of the characteristics of System 1, the experiential or intuitive system, is its affective basis. Although analysis (System 2) is certainly important in many decision-making circumstances, reliance on affect and emotion is generally a quicker, easier, and more efficient way to navigate in a complex, uncertain, and sometimes dangerous world. Many theorists have given affect a direct and primary role in motivating behavior.

Underlying the role of affect in the experiential system is the importance of images, to which positive or negative feelings become attached. Images in this system include not only visual images, important as these may be, but words, sounds, smells, memories, and products of our imagination.

Kahneman (2003) noted that one of the functions of System 2 is to monitor the quality of the intuitive impressions formed by System 1. Kahneman and Frederick (2002) suggested that this monitoring is typically rather lax and allows many intuitive judgments to be expressed in behavior, including some that are erroneous. This point has important implications that will be discussed later.

In addition to positive and negative affect, more nuanced feelings such as empathy, sympathy, compassion, and sadness have been found to be critical for motivating people to help others (Coke, Batson, and McDavis, 1978; Dickert and Slovic, 2009; Eisenberg and Miller, 1987). As Batson (1990) put it, "considerable research suggests that we are more likely to help someone in need when we 'feel for' that person." (p. 339).

A particularly important psychological insight comes from Haidt (2001, 2007; see also Van Berkum et al., 2009), who argued that moral intuitions (akin to System 1) precede moral judgments. Specifically, he asserted that

> moral intuition can be defined as the sudden appearance in consciousness of a moral judgment, including an affective valence (good-bad, like-dislike) without any conscious awareness of having gone through steps of searching, weighing evidence, or inferring a conclusion. Moral intuition is therefore . . . akin to aesthetic judgment. One sees or hears about a social event and one instantly feels approval or disapproval. (p. 818)

In other words, feelings associated with moral intuition usually dominate moral judgment, unless we make an effort to use judgment to critique and, if necessary, override intuition. Not that our moral intuitions are not, in many cases, sophisticated and accurate. They are much like human visual perception in this regard—equipped with shortcuts that most of the time serve us well but occasionally lead us seriously astray (Kahneman, 2003). Indeed, like perception, which is subject under certain conditions to visual illusions, our moral intuitions can be very misguided. We shall demonstrate this in the following sections

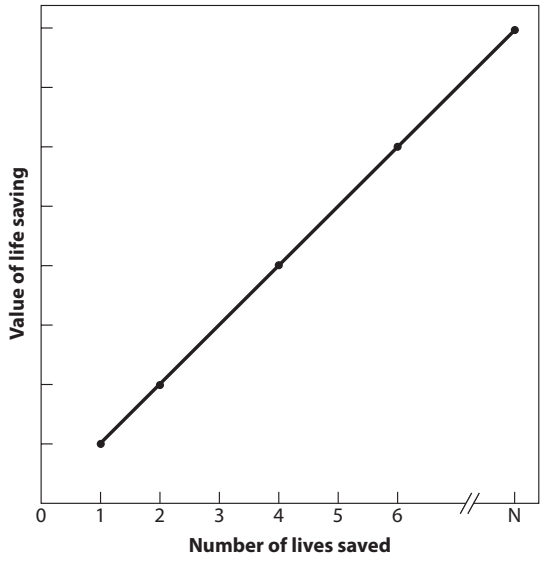
and argue that, in particular, our intuitions fail us in the face of genocide and mass atrocities. That failure points to the need to create laws and institutions that are designed to stimulate reasoned analysis and that can help us overcome the deficiencies in our ability to *feel* the need to act.

Affect, Analysis, and the Value of Human Lives

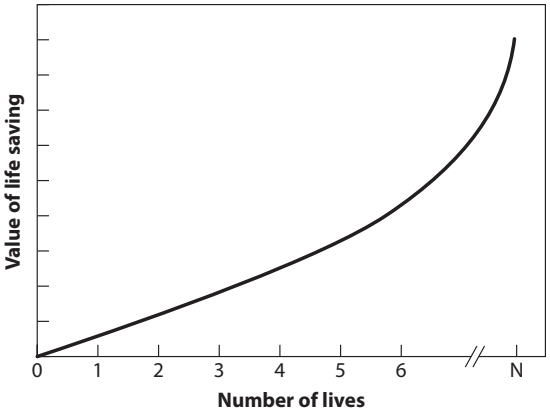
How *should* we value the saving of human lives? A System 2 answer would look to basic principles or fundamental values for guidance. For example, Article 1 of the United Nations Universal Declaration of Human Rights asserts, "All human beings are born free and equal in dignity and rights." We might infer from this that every human life is of equal value. If so, the value of saving *N* lives is *N* times the value of saving one life, as represented by the linear function in figure 7.1.

An argument can also be made for judging large losses of life to be disproportionately more serious because they threaten the social fabric and viability of a group or community, as in genocide (fig. 7.2). Debate can be had at the margins over whether governments have a duty to give more weight to the lives of their own people, but something approximating the equality of human lives is rather uncontroversial.

How *do* we actually value human lives? Research provides evidence in support of two descriptive models, which are linked to affect and intuitive, System 1 thinking, that reflect values for lifesaving that are profoundly different from the normative models shown in figures 7.1 and 7.2. Both of these descriptive models demonstrate responses that are insensitive to large losses of human life and consistent with apathy toward genocide.



7.1. A normative model for valuing the saving of human lives. Every human life is of equal value.



7.2. Another normative model: large losses threaten the viability of the group or society.

The Psychophysical Model

Affect is a remarkable mechanism that enabled humans to survive the long course of evolution. Before there were sophisticated analytic tools such as probability theory, scientific risk assessment, and cost/benefit calculus, humans used their senses, honed by experience, to determine whether the animal lurking in the bushes was safe to approach or the murky water in the pond was safe to drink. Simply put, System 1 thinking evolved to protect individuals and their small family and community groups from present, visible, immediate dangers. This affective system, however, did not evolve to help us respond to distant, mass murder. As a result, System 1 thinking responds to large-scale atrocities in ways that System 2 deliberation, if activated, finds reprehensible.

Fundamental qualities of human behavior are, of course, recognized by others beside scientists. The American writer Annie Dillard (1999) cleverly demonstrated the limitation of our affective system as she sought to help us understand the humanity of the Chinese nation: "There are 1,198,500,000 people alive now in China. To get a *feel* for what this *means*, simply take yourself—in all your singularity, importance, complexity, and love—and multiply by 1,198,500,000. See? Nothing to it" (p. 47, italics added).

We quickly recognize that Dillard was joking when she asserted "nothing to it." We know, as she did, that

we are incapable of *feeling* the humanity behind the number 1,198,500,000. The circuitry in our brain is not up to this task. This same incapacity was purportedly echoed by the Nobel Prize–winning biochemist Albert Szent-Györgyi as he struggled to comprehend the possible consequences of nuclear war: "I am deeply moved if I see one man suffering and would risk my life for him. Then I talk impersonally about the possible pulverization of our big cities, with a hundred million dead. I am unable to multiply one man's suffering by a hundred million."

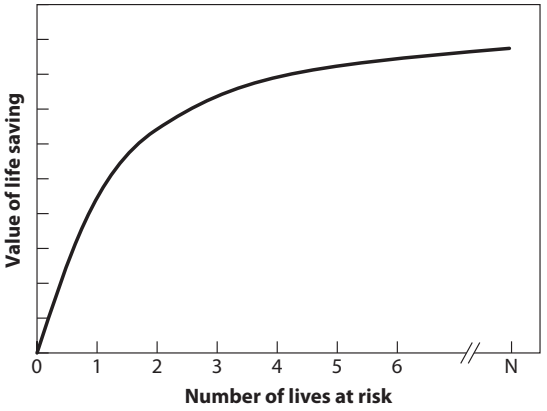
There is considerable evidence that our affective responses and the resulting value we place on saving human lives may follow the same sort of "psychophysical function" that characterizes our diminished sensitivity to a wide range of perceptual and cognitive entities—brightness, loudness, heaviness, and money—as their underlying magnitudes increase.

What psychological principle lies behind this insensitivity? In the nineteenth century, E. H. Weber (1834) and Gustav Fechner (1860/1912) discovered a fundamental psychophysical principle that describes how we perceive changes in our environment. They found that people's ability to detect changes in a physical stimulus rapidly decreases as the magnitude of the stimulus increases. What is known today as Weber's law states that in order for a change in a stimulus to become *just noticeable*, a fixed percentage must be added. Thus, perceived difference is a relative matter. To a small stimulus, only a small amount must be added to be noticeable. To a large stimulus, a large amount must be added. Fechner proposed a logarithmic law to model this nonlinear growth of sensation. Numerous empirical studies by S. S. Stevens (1975) have demonstrated that the growth of sensory magnitude (ψ) is best fit by a power function of the stimulus magnitude φ,

$$\psi = k\varphi\beta,$$

where the exponent β is typically less than one for measurements of phenomena such as loudness, brightness, and even the value of money (Galanter, 1962). For example, if the exponent is 0.5, as it is in some studies of perceived brightness, a light that is four times the intensity of another light will be judged only twice as bright.

Remarkably, the way that numbers are represented mentally may also follow the psychophysical function. Dehaene (1997) described a simple experiment in which people are asked to indicate which of two numbers is larger: 9 or 8? 2 or 1? Everyone gets the answers right, but it takes more time to identify 9 as larger than 8 than to indicate 2 is larger than 1. From experiments such as this, Dehaene concluded that "our brain represents quantities in a fashion not unlike the logarithmic scale on a slide rule, where equal space is allocated to the interval between 1 and 2, 2 and 4, or between 4 and 8" (p. 76). Numbers 8 and 9 thus seem closer together or more similar than 1 and 2.



7.3. A psychophysical model describing how the saving of human lives may actually be valued.

Our cognitive and perceptual systems seem designed to sensitize us to small changes in our environment, possibly at the expense of making us less able to detect and respond to large changes. As the psychophysical research indicates, constant increases in the physical magnitude of a stimulus typically evoke smaller and smaller changes in response. Applying this principle to the valuing of human life suggests that a form of *psychophysical numbing* may result from our inability to appreciate losses of life as they become larger (fig. 7.3). The function in figure 7.3 represents a value structure in which the importance of saving one life is great when it is the first, or only, life saved but diminishes marginally as the total number of lives saved increases. Thus, psychologically, the importance of saving one life is diminished against the background of a larger threat—we will likely not "feel" much difference, nor value the difference, between saving 87 lives and saving 88.

Kahneman and Tversky (1979) incorporated this psychophysical principle of decreasing sensitivity into *prospect theory*, a descriptive account of decision making under uncertainty. A major element of prospect theory is the value function, which relates subjective value to actual gains or losses. When applied to human lives, the value function implies that the subjective value of saving a specific number of lives is greater for a smaller tragedy than for a larger one.

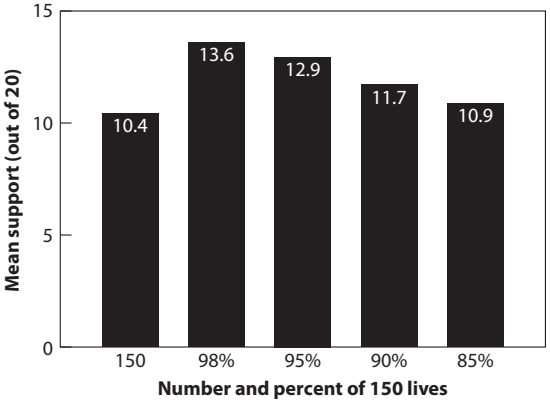
Fetherstonhaugh et al. (1997) demonstrated this potential for diminished sensitivity to the value of life—that is, *psychophysical numbing*—in the context

of evaluating people's willingness to fund various lifesaving interventions. In a study involving a hypothetical grant-funding agency, respondents were asked to indicate the number of lives a medical research institute would have to save to merit receipt of a $10 million grant. Nearly two-thirds of the respondents raised their minimum benefit requirements to warrant funding when there was a larger at-risk population, with a median value of 9,000 lives needing to be saved when 15,000 were at risk, compared to a median of 100,000 lives needing to be saved out of 290,000 at risk. By implication, respondents saw saving 9,000 lives in the smaller population as more valuable than saving ten times as many lives in the larger population.

Other studies in the domain of lifesaving interventions have documented similar psychophysical numbing or proportional reasoning effects (Baron, 1997; Bartels and Burnett, 2006; Fetherstonhaugh et al., 1997; Friedrich et al., 1999; Jenni and Loewenstein, 1997; Ubel, Baron, and Asch, 2001). For example, Fetherstonhaugh et al. (1997) also found that people were less willing to send aid that would save 4,500 lives in Rwandan refugee camps as the size of the camps' at-risk population increased. Friedrich et al. (1999) found that people required more lives to be saved to justify mandatory antilock brakes on new cars when the alleged size of the at-risk pool (annual braking-related deaths) increased.

These diverse studies of lifesaving demonstrate that the *proportion* of lives saved often carries more weight than the *number* of lives saved when people evaluate interventions. Thus, extrapolating from Fetherstonhaugh et al., one would expect that, in separate evaluations, there would be more support for saving 80% of 100 lives at risk than for saving 20% of 1,000 lives at risk. This is consistent with an affective (System 1) account, in which the number of lives saved conveys little affect but the proportion saved carries much feeling: 80% is clearly "good" and 20% is "poor."

Slovic et al. (2004), drawing upon the finding that proportions appear to convey more feeling than do numbers of lives, predicted (and found) that college students, in a between-groups design, would more strongly support an airport-safety measure expected to save 98% of 150 lives at risk than a measure expected to save 150 lives. Saving 150 lives is diffusely good, and therefore somewhat hard to evaluate, whereas saving 98% of something is clearly very good because it is so close to the upper bound on the percentage scale and hence is highly weighted in the support judgment. Subsequent reduction of the percentage of 150 lives that would be saved to 95%, 90%, and 85% led to reduced support for the safety measure, but each of these percentage conditions still garnered a higher mean level of support than did the "save 150 lives" condition (fig. 7.4).



7.4. Airport safety study: saving a percentage of 150 lives receives higher support ratings than does saving 150 lives.

Note: Bars describe mean responses to the question, How much would you support the proposed measure to purchase new equipment? The response scale ranged from 0 (would not support at all) to 20 (very strong support).

(Data derived from table 23.4, page 408, from chapter 23, "The Affect Heuristic," Slovic et al., *Heuristics and Biases: The Psychology of Intuitive Judgment*, edited by Thomas Gilovich, Dale Griffin and Daniel Kahneman. Copyright © 2002, Cambridge University Press. Adapted with permission.)

This research on psychophysical numbing is important because it demonstrates that the feelings necessary for motivating lifesaving actions are not congruent with the normative models in figures 7.1 and 7.2. The nonlinearity displayed in figure 7.3 is consistent with the disregard of incremental loss of life against a background of a large tragedy. However, it does not fully explain apathy toward genocide, because it implies that the response to initial loss of life will be strong and maintained, albeit with diminished sensitivity, as the losses increase. Evidence for a second descriptive model, better suited to explain apathy toward genocide, follows.

Numbers and Numbness: Images and Feeling

Psychological theories and data confirm what keen observers of human behavior have long known. Numerical representations of human lives do not necessarily convey the importance of those lives. All too often the numbers represent dry statistics, "human beings with the tears dried off," that lack feeling and fail to motivate action (Slovic and Slovic, 2004).

How can we impart the feelings that are needed for rational action? Attempts to do this typically involve highlighting the images that lie beneath the

7.5. Flags depicting American and Iraqi war dead.
(Reprinted from "Affect, Moral Intuition, and Risk," Slovic, Paul, and Daniel Västfjäll, *Psychological Inquiry, 21*(4), 1.10.2010. Reprinted by permission of Taylor & Francis Ltd.)

numbers. For example, organizers of a rally designed to get Congress to do something about 38,000 deaths a year from handguns piled 38,000 pairs of shoes in a mound in front of the Capitol (Associated Press, 1994). Students at a middle school in Tennessee, struggling to comprehend the magnitude of the Holocaust, collected six million paper clips as a centerpiece for a memorial (Schroeder and Schroeder-Hildebrand, 2004). Flags were "planted" on the lawn of the University of Oregon campus to represent the thousands of American and Iraqi war dead (fig. 7.5).

When it comes to eliciting compassion, the identified individual victim, with a face and a name, has no peer. Psychological experiments demonstrate this clearly, but we all know it as well from personal experience and media coverage of heroic efforts to save individual lives. The world watched tensely as rescuers worked for several days to rescue 18-month-old Jessica McClure, who had fallen 22 feet into a narrow abandoned well shaft. Charities such as Save the Children have long recognized that it is better to endow a donor with a single, named child to support than to ask for contributions to the bigger cause.

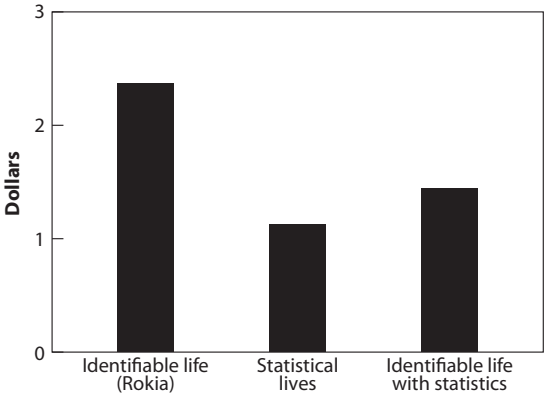
Even Adolf Eichmann, complicit in the murder of millions of Jews during the Holocaust, exhibited an emotional connection to one of his victims after being interrogated by the victim's son for hundreds of hours during his 1961 trial in Israel. When the interrogator, Captain Avner Less, reveals to Eichmann that his father had been deported to Auschwitz by Eichmann's headquarters, Eichmann cried out, "But that's horrible, Herr Captain! That's horrible!" (von Lang, 1983, p. ix).

But the face need not even be human to motivate powerful intervention. A dog stranded aboard a tanker adrift in the Pacific was the subject of one of the most costly animal rescue efforts ever (Vedantam, 2010). Hearing this, the columnist Nicholas Kristof (2007) recalled cynically that a single hawk, Pale Male, evicted from his nest in Manhattan, aroused more indignation than two million homeless Sudanese. He observed that what was needed to galvanize the American public and their leaders to respond to the genocide in Darfur was a suffering puppy with big eyes and floppy ears: "If President Bush and the global public alike are unmoved by the slaughter of hundreds of thousands of fellow humans, maybe our last, best hope is that we can be galvanized by a puppy in distress."

The Collapse of Compassion

In recent years, vivid images of natural disasters in Southeast Asia, on the American Gulf Coast, and in Haiti and stories of individual victims there brought to us through relentless, courageous, and intimate news coverage unleashed an outpouring of compassion and humanitarian aid from all over the world. Perhaps there is hope here that vivid, personalized media coverage featuring victims of genocide could motivate intervention to prevent mass murder and genocide.

Perhaps. Research demonstrates that people are much more willing to aid identified individuals than unidentified or statistical victims (Jenni and Loewenstein, 1997; Kogut and Ritov, 2005a; Schelling, 1968; Small and Loewenstein, 2003, 2005). But a cautionary note comes from a study by Small, Loewenstein, and Slovic (2007), who gave people leaving a psychological experiment the opportunity to contribute up to $5 of their earnings to Save the Children. In one condition, respondents were asked to donate money to feed an identified victim, a seven-year-old African girl named Rokia. They contributed more than twice the amount given by a second group asked to donate to the same organization working to save millions of Africans from hunger (fig. 7.6). Respondents in a



7.6. Mean donations.
(Redrawn from Small et al., 2007)

third group were asked to donate to Rokia but were also shown the larger statistical problem (millions in need) shown to the second group. Unfortunately, coupling the statistical realities with Rokia's story significantly *reduced* the contributions to Rokia. It may be that the presence of statistics reduced the attention to Rokia essential for establishing the emotional connection necessary to motivate donations.

Alternatively, the recognition of the millions not being helped by one's donation may have produced negative affect that inhibited the response.

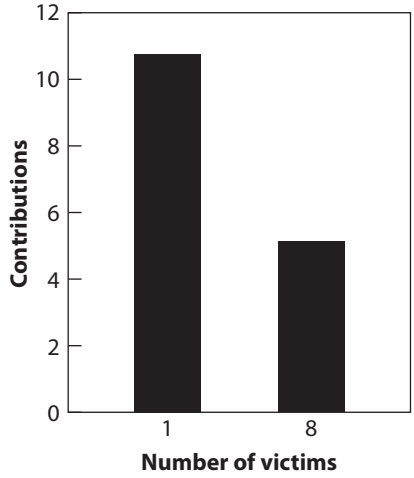
A follow-up experiment by Small, Loewenstein, and Slovic (2007) provided additional evidence for the importance of feelings. Before being given the opportunity to donate, participants were either primed to feel ("Describe your feelings when you hear the word *baby*," and similar items) or to do simple arithmetic calculations. Priming analytic thinking (calculation) reduced donations to the identifiable victim (Rokia) relative to the feeling prime. Yet the two primes had no distinct effect on statistical victims, which is symptomatic of the difficulty in generating feelings for such victims.

Annie Dillard read in her newspaper the headline "Head Spinning Numbers Cause Mind to Go Slack." She struggled to think straight about the great losses that the world ignores: "More than two million children die a year from diarrhea and eight hundred thousand from measles. Do we blink? Stalin starved seven million Ukrainians in one year, Pol Pot killed one million Cambodians." She writes of "compassion fatigue" and asks, "At what number do other individuals blur for me?" (Dillard, 1999, pp. 130–131).

An answer to Dillard's question is beginning to emerge from behavioral research. Studies by Hamilton and Sherman (1996) and Susskind et al. (1999) found that a single individual, unlike a group, is viewed as a psychologically coherent unit. This leads to more extensive processing of information and stronger impressions about individuals than about groups. Consistent with this, Kogut and Ritov (2005a, 2005b) found that people tend to feel more distress and compassion when considering an identified single victim than when considering a group of victims, even if identified.

Specifically, Kogut and Ritov asked participants to contribute to a costly lifesaving treatment needed by a sick child or a group of eight sick children. The target amount needed to save the child (children) was the same in both conditions. All contributions were actually donated to children in need of cancer treatment. In addition, participants rated their feelings of distress (feeling worried, upset, and sad) toward the sick child (children).

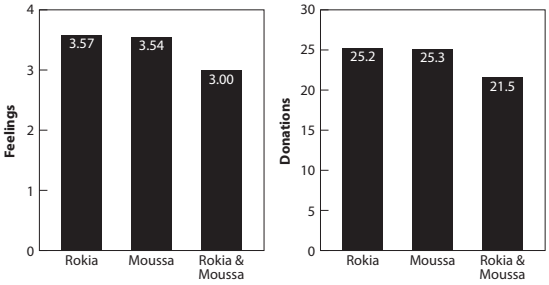
The mean contributions are shown in figure 7.7. Contributions to the individuals in the group, as individuals, were far greater than were contributions to the entire group. Ratings of distress were also higher in the individual condition. Kogut and Ritov concluded that the greater donations to the single victim most likely stem from the stronger emotions evoked by such victims.



7.7. Mean contributions to individuals and their group. (Data from Kogut and Ritov, 2005b)

Västfjäll, Peters, and Slovic (2010) decided to test whether the effect found by Kogut and Ritov would occur as well for donations to two starving children. Following the protocol designed by Small, Loewenstein, and Slovic (2007), they gave one group of Swedish students the opportunity to contribute their earnings from another experiment to Save the Children to aid Rokia, whose plight was described as in the study by Small and coworkers. A second group was offered the opportunity to contribute their earnings to Save the Children to aid Moussa, a seven-year-old boy from Africa who was similarly described as in need of food aid. A third group was shown the vignettes and photos of Rokia and Moussa and was told that any donation would go to both of them, Rokia *and* Moussa. The donations were real and were sent to Save the Children. Participants also rated their feelings about donating on a scale of 1 (negative) to 5 (positive). Affect was found to be least positive in the combined condition, and donations were smaller in that condition (fig. 7.8). In the individual-child conditions, the size of the donation made was strongly correlated with rated feelings (r = .52 for Rokia; r = .52 for Moussa). However this correlation was much reduced (r = .19) in the combined condition.

As unsettling as is the valuation of lifesaving portrayed by the psychophysical model, the studies just described suggest an even more disturbing psychological tendency. Our capacity to feel is limited. To



7.8. Feelings and donations decline at $N = 2$! Mean affect ratings (*left*) and mean donations (*right*) for individuals and their combination.
(Adapted from Västfjäll et al., 2010)

7.9. A model depicting psychic numbing—the collapse of compassion—when valuing the saving of lives.

the extent that the valuation of lifesaving depends on feelings driven by attention or imagery, it might follow the function shown in figure 7.9, where the emotion or affective feeling is greatest at $N = 1$ but begins to decline at $N = 2$ and collapses at some higher value of N that becomes simply "a statistic." In other words, returning to Dillard's worry about compassion fatigue, perhaps the "blurring" of individuals begins at two! Whereas Lifton (1967) coined the term *psychic numbing* to describe the "turning off" of feeling that enabled rescue workers to function during the horrific aftermath of the Hiroshima bombing, figure 7.9 depicts a form of numbing that is not beneficial. Rather, it leads to apathy and inaction, consistent with what is seen repeatedly in response to mass murder and genocide.

The Failure of Moral Intuition

Thoughtful deliberation takes effort. Fortunately evolution has equipped us with sophisticated cognitive and perceptual mechanisms that can guide us through our daily lives efficiently with minimal need for "deep thinking." We have referred to these mechanisms as System 1.

Consider, for example, how we deal with risk. Long before we had invented probability theory, risk assessment, and decision analysis, there was intuition, instinct, and gut feeling, aided by experience, to tell us whether an animal was safe to approach or the water was safe to drink. As life became more complex and we gained more control over our environment, analytic ways of thinking, known as System 2, evolved to boost the rationality of our experiential reactions. Beyond the question of how water looks and tastes, we now look to toxicology and analytic chemistry to tell us whether the water is safe to drink (Slovic et al., 2004). But we can still use our feelings as well, an easier path.

As with risk, the natural and easy way to deal with moral issues is to rely on our intuitions: "How bad is it?" Well, how bad does it feel? We can also apply reason and logical analysis to determine right and wrong, as our legal system attempts to do. But moral intuition comes first and usually dominates moral judgment unless we make an effort to use judgment to critique and, if necessary, override our intuitive feelings (Haidt, 2001, 2007).

Unfortunately, moral intuition fails us in the face of genocide and other disasters that threaten human lives and the environment on a large scale. As powerful as System 1 is, when infused with vivid experiential stimulation (witness the moral outrage triggered by the photos of abuse at the Abu Ghraib prison in Iraq), it has a darker side. We cannot trust it. It depends upon attention and feelings that may be hard to arouse and sustain over time for large numbers of victims, not to speak of numbers as small as two. Left to its own devices, moral intuition will likely favor individual victims and sensational stories that are closer to home and easier to imagine. It will be distracted by images that produce strong, though erroneous, feelings, like percentages as opposed to actual numbers. Our sizable capacity to care for others may also be overridden by more pressing personal interests. Compassion for others has been characterized by Batson et al. (1983) as "a fragile flower, easily crushed by self concern" (p. 718). Faced with genocide and other mass tragedies, we cannot rely on our moral intuitions alone to guide us to act properly.

Philosophers such as Peter Singer (2007) and Peter Unger (1996), employing very different methods than psychologists, came to much the same conclusions about the unreliability of moral intuitions. Unger, after leading his readers through fifty ingenious thought experiments, urged them and us to think harder to overcome the morally questionable

appearances promoted by our intuitive responses. These intuitions, he argued, lead us to act in ways that are inconsistent with our true "Values," that is, the values we would hold after more careful deliberation: "Folks' intuitive moral responses to specific cases derive from sources far removed from our Values and, so, they fail to reflect these Values, often even pointing in the opposite direction" (p. 11).

Greene (2008), drawing on data from psychology and neuroscience as well as philosophy, attempted to explain the problems with intuitions in terms of the morally irrelevant evolutionary factors that shaped these intuitions. Thus we say it is wrong to abandon a drowning child in a shallow pond but okay to ignore the needs of millions of starving children abroad. The former pushes our emotional buttons whereas the latter do not. And this may be because we evolved in an environment in which we lived in small groups and developed immediate, emotionally based intuitive responses to the needs and transgressions of others. There was little or no interaction with faraway strangers.

Implications for International Law and Policy

Clearly there are many serious obstacles to consistent, meaningful intervention to prevent genocide and similarly grave abuses. In addition to the more obvious political, material, and logistical impediments, the international community must overcome the psychological constraints described here. Indeed, the cognitive limitations we identify make it much more difficult to mobilize global public sentiment in the way necessary to overcome the more obvious material and logistical constraints. The question is whether and how international law and institutions might be reformed to account for these cognitive limitations. In this section, we will canvass several implications of this research for the law and policy of atrocity prevention.

Although we have emphasized the implications of this research for the problem of genocide, much of the psychological research obviously applies to violations involving large numbers of victims in general. The data are not limited to genocide. The research provides insight into the ways in which individuals react to mass human rights abuses such as widespread arbitrary detentions and denial of a population's right to food. As a consequence, the lessons that are relevant to policy makers and practitioners relate broadly to the field of human rights.

Several of the following proposals are ambitious—especially those involving a change to the use-of-force regime—and that ambition raises questions about their political viability. But there are several factors that may increase their viability. First, attempts to identify and reduce psychic numbing among electorates may produce a political opportunity for institutional change—that is, to the extent that psychic numbing exists and is masking a preference for antigenocide action, unmasking that preference may produce powerful political will. Second, political actors themselves may be more willing to embrace these various reforms if the changes are not intended to overcome political interests, but instead to overcome cognitive failures. The psychological research shows a collapse of rational calculation and evaluation that causes us to artificially devalue human life. Indeed, in some circumstances, the more widespread and systematic the violation, the weaker the reaction. At bottom, the need for reform should be grounded in an understanding that cognitive deficiencies can prevent actors from realizing a preference for stopping mass human rights violations—*even when doing so would serve their overall values and interests.*

Appreciation of the failure of moral intuition should inform the development of new legal rules and institutional arrangements concerned with atrocity prevention and human rights more generally. Indeed, it may only be laws and institutions that can keep us on course, forcing us to pursue the hard measures needed to combat massive human rights abuses when our attention strays and our feelings lull us into complacency. We accordingly propose several institutional designs to improve international decision making in this arena. We discuss several strategies (1) to insulate institutions from the effects of psychic numbing; (2) to remove or restrict institutional features that foster psychic numbing; (3) to promote System 2 deliberation directly; and (4) to employ System 1 to channel actors toward System 2 processes.

Insulate Institutions from the Effects of Psychic Numbing

One approach is to insulate decision-making processes from the adverse psychological effects that we have identified. For example, policy makers might design institutions to be less susceptible to psychic numbing or to operate despite the psychological effects on actors within the institution.

CONSTRUCT DEFAULT RULES AND PRECOMMITMENT DEVICES

The international regime could construct precommitment enforcement strategies to deal with genocide and other human rights atrocities of similar scale. Consider a few options: the UN Security Council could preauthorize, subject perhaps to an *ex post*

council override, the use of force in any situation in which atrocities reach a certain scale. Another possibility is that the Security Council could order (rather than authorize) all member states to take coercive action once the commission of atrocities reached a certain level. Alternatively, states could conclude a treaty in which state parties would preinvite foreign intervention and/or UN peacekeepers in the event that genocide occurs on their own territory.

Similarly, the psychological evidence provides a powerful reason for supporting the responsibility to protect, an emerging doctrine that shifts from a *right* of states to a *duty* of states to intervene in another country to stop an atrocity (Wheeler, 2005). The novelty of the responsibility to protect is that states are under an affirmative obligation—not just a license—to intervene once the Security Council has authorized such action. The psychological findings provide an independent and unique reason to place pressure on states in the form of this legal responsibility. The starting point should favor intervention (at the very least when the Security Council has determined force is appropriate).

Other precommitment strategies could be implemented to insulate institutions from the effects of psychic numbing with respect to human rights more generally. Aside from the use-of-force regime, multilateral organizations could preauthorize economic sanctions on the part of their member states. Nations could pass domestic legislation that triggers such sanctions or automatically increases foreign aid in the event of a humanitarian catastrophe (and could perhaps require repeal of such aid by a supermajority). States could preauthorize UN Special Rapporteurs to visit their country in the event of mass human rights violations. In all these instances, multilateral bodies, foreign countries, and the affected nation might be ill equipped—without the assistance of a precommitment device—to confront a situation after deaths and deprivations begin to mount.

Questions about whether and how to intervene in ongoing conflicts—militarily, economically, etc.—tend to occupy the field of the genocide-response debate, and one appealing feature of the psychic numbing literature is that it may offer a simple metric for determining *when* to intervene. Say, for example, that valuations of life begin to drop off significantly after 10 deaths. At 10 deaths, a preauthorized UN investigation would automatically be triggered (implementing new reporting methods, as discussed below); at 100 deaths, that investigatory body would immediately acquire certain authorities. These lockstep provisions can be justified on the grounds that any more-subjective metric raises the risk of psychic numbing. If such a system could be implemented, it could limit the opportunity for genocidaire states to stall international intervention under the guise of diplomatic debate.

EMPHASIZE EARLY WARNING AND PREVENTIVE ACTION

Another approach is to act before psychic numbing sets in. Apart from the fact that prevention is in many ways easier, less costly, and less difficult than intervention (Hamburg, 2008), reaction strategies must necessarily overcome the psychic numbing generated by the instant crisis. This insight recommends a range of law and policy options including more vigorous international monitoring or intervention in situations likely to generate wide-scale atrocities (e.g., civil wars, military coups, etc.) or even "anticipatory" humanitarian intervention (Richter and Stanton, n.d.). It recommends establishing a general, preventive disclosure mechanism to preclude trafficking in resources that are at risk for funding human rights abuses, as a recent U.S. law attempts to do for conflict minerals in Congo. It also calls for greater financial and political support for criminal trials—if that instrument can be expected to deter future violations or to help halt cycles of violence. Prevention and reaction strategies need not be mutually exclusive. Early-warning systems could feasibly be triggered after low numbers of deaths, but such triggers need not be more than an investigatory panel to assess the risks of the current situation, both for future harm and also the risk of psychic numbing as the situation develops. That is, early-warning systems can emphasize both prevention and preparation—in this case preparation for numbing effects.

EMPOWER INSTITUTIONS AND ACTORS LESS LIKELY TO SUCCUMB TO PSYCHIC NUMBING

Psychological research also provides good reasons to support a form of subsidiarity within the humanitarian rights and use-of-force regime. Regional and local actors who are closer to the situation are more likely to overcome System 1 limitations in comprehending the gravity of an atrocity. Accordingly, international law might provide regional organizations (e.g., the Economic Community of West African States, the African Union) greater leeway to use force to stop genocide before or even without Security Council action. The objective here is to create a one-way ratchet—providing more proximate and local actors an option to intervene without complete international backing. The design would not work the other way; that is, to provide regional actors the authority to bar outside intervention by the international community.

Regional actors could also be empowered in intergovernmental settings involving enforcement measures

that do not entail the use of force. Examples of such enforcement measures include formal resolutions condemning a state for extremely poor human rights conditions, the creation of a special rapporteur to monitor the country, the ouster of a state from an intergovernmental organization, and the imposition of economic sanctions. Voting rules could be fashioned whereby such measures would be adopted *either* if a majority of state parties approves *or* if a supermajority of states from the relevant region approve. For example, the imposition of sanctions against Zimbabwe could be approved either (1) by a majority of all state parties to an international organization or (2) by approval of three-quarters of the African states even if a majority of the whole does not agree. Once again, these devices are intended to function as a one-way ratchet. Such a design principle would be important because of other political and psychological reasons that regional actors may otherwise protect their neighbors from enforcement actions.

Outside monitoring and independent international review are key components of the international regime. The foregoing discussion suggests that, in fact, outside reviewers may be more susceptible to numbing effects. Responding by empowering local actors to conduct investigations may not solve the numbing problem but may instead replace it with a neutrality problem if the local actors are less likely than their international counterparts to be impartial observers. One potential solution would be to turn to intermediate actors—regional bodies or hybrid local-international bodies. Another would be to train, to the extent possible, the relevant rapporteurs to recognize and counter the risks of psychic numbing. But however the problem is addressed, institutional capacities must be assessed and—although it is not currently recognized as such—psychic numbing is a relevant factor to consider in making this assessment.

Remove or Restrict Institutional Features That Foster Psychic Numbing

CHANGE THE METHOD AND CONTENT OF HUMAN RIGHTS REPORTING

By challenging the assumption that information makes positive change more likely, the research presented in this chapter calls into question one of the strategic pillars of human rights advocacy. Documentation—including the presentation of data showing mass and systematic violations—is often thought to raise awareness. Efforts by international organizations to document mass human rights violations typically focus on the widespread nature of violations rather than on narratives or other information about the individuals who have been harmed. Statistics prevail over stories. A good example of this is the Darfur Atrocities Documentation Project (Totten, 2006), which compiled a database of over 10,000 eye-witnessed incidents but reported mostly the percentages of different types of abuses.

International legal procedures amplify the problem. First, consider the strict page limitations that exist for reports to the UN Human Rights Council. These page constraints apply to reports by nongovernmental organizations as well as to those by UN human rights officials. As a result, the authors of the reports condense information into compact pieces of data and are unable to delve deeply into descriptions of individuals' lives. Under these pressures, statistics are also considered an efficient method for conveying information. Second, in official settings little opportunity exists for conveying information in the form of visual media. Third, important international legal forums impose either an expressed or implicit requirement that violations meet a quantitative threshold (UN Human Rights Council's 1503 Complaints Procedure), which incentivizes advocates to frame their appeal through the representation of large numbers of cases. It is not difficult to conceive of innovations to repair these problems. Procedural and substantive requirements could be softened or exceptions could be made to expand the forms of information conveyance.

RECONSIDER HUMAN RIGHTS INDICATORS

Many now call for the use of quantitative indicators in global governance (e.g., measures of good governance by the World Bank; see, e.g., Davis, Kingsbury, and Merry, 2010). The psychological research documented here suggests that significant perverse effects may result from the production, collection, and circulation of quantitative human rights indicators. Actors involved in these processes may become desensitized to human rights violations, and such processes often involve some of the most important actors within government and civil society. These effects may not be a sufficient basis to abandon or restrict indicators; however, in the emerging debate about their utility, these risks should be carefully considered.

Nevertheless, indicators can prove invaluable for monitoring and responding to psychic numbing. First, indicators can provide a valuable tool for tracking the likelihood of numbing effects—the larger the numbers involved, the greater the risks. Second, we can acknowledge the possibility that indicators might induce numbing without abandoning their use. Instead, we must be mindful of the difference between the collection of data and its final presented form. Data collection and data reporting could be done

by different agencies, and data collectors should be guarded against numbing effects and also trained to look for stories that can serve to illustrate the significance of a given atrocity.

RECONSIDER SUBSTANTIVE ELEMENTS OF HUMAN RIGHTS LAW

Even the substantive law of genocide might be considered problematic since it conceptualizes genocide as a collective or group injury rather than as harm to individuals. As a result of the legal definition, the discourse surrounding the presentation of grievances may focus too extensively on the group-based harms. In this light, it is instructive to reflect on the characterization by a Holocaust survivor, Abel Hertzberg: "There were not six million Jews murdered: there was one murder, six million times" (U.S. Holocaust Memorial Museum, 2005).

The definition of *crimes against humanity* raises a similar concern. Generally defined as a "widespread and systematic" attack against a civilian population, the elements of the crime might emphasize the representation of aggregate numbers rather than of individual cases. The particular definition of crimes against humanity in the UN Statute for the Rwanda Tribunal includes an unusual requirement: that the attack be directed against a "civilian population on national, political, ethnic, racial or religious grounds." That definition (which was altered in the treaty for the International Criminal Court) is subject to some of the same concerns as the group-based focus of genocide.

Employ System 1 to Activate and Support System 2 Processes

Despite the limitations of System 1 noted above, we should nevertheless attempt to bolster it, at least so it can motivate support for efforts based on System 2. Such attempts should capitalize on the findings described earlier demonstrating that we care most about aiding individual people in need, even more so when we can attach a name and a face to them.

AFFECTIVE IMAGERY

The data in this chapter present a striking irony: in an effort to emphasize objective facts, the human rights regime risks losing its ability to connect with sympathizers on a human level. To be sure, we do not advocate wholesale abandonment of current reporting mechanisms or the exclusive adoption of emotion-laden stories. After all, the goal of overcoming psychic numbing is to better calibrate our interventions to the scale of the atrocities that we face. But there is ample room for the future of human rights reporting to exhibit mixed methodologies.

The increasing availability of mixed media may help in this regard. As people post visceral digital content depicting human rights abuses, audiences may exhibit responses that otherwise had been masked by numbing effects. In April 2010, the website Wikileaks posted video of U.S. soldiers firing indiscriminately upon civilians in Iraq, creating a media and political uproar. Dozens of news reports had already reported on the problem of indiscriminate targeting, none of which garnered the same attention as the online video. The same phenomenon can be said of the Abu Ghraib prisoner abuse scandal—during the entire U.S. occupation of Iraq, nothing created the same backlash as the release of the photos of prisoner mistreatment, despite several reports that, although less colorful, suggested much more violent and more widespread practices.

Thus, one possibility is to infuse human rights reporting with powerful affective imagery, such as that associated with Hurricane Katrina, the Southeast Asian tsunami, and the earthquake in Haiti. This would require pressure on the media to report the slaughter of innocent people aggressively and vividly. Another way to engage our experiential system would be to bring people from abused populations into our communities and our homes to tell their stories.

Above we discussed the disadvantages of reports that focus on the numbers of violations. While it is obviously necessary to document the scope of such atrocities, neglecting the stories of individuals certainly contributes to numbing. Human rights advocates should reorient the documentation and reporting of abuses to prompt System 1 thinking. In some cases, in-depth narratives and visual personal stories describing the predicament of individual victims should be emphasized instead of more abstract descriptions of the scale of abuses—that is, stories over statistics.

At the same time, scale and systematicity presumably remain important for calibrating the appropriate response to any human rights problem. As a consequence, human rights documentation should not abandon the reporting of scale and system-level effects. The central challenge of applying the psychological research to human rights advocacy is identifying when or how many "statistics" and when or how much "storytelling" should be employed in the documentation and reporting of abuses. Arresting visual displays (such as that shown in fig. 7.5) and photographs of victims and atrocities should be included in the reporting and in the publicly distributed

information presented by human rights advocates. Indeed, the future success of the human rights movement requires training not only advocates skilled in documenting large numbers of cases and professionals skilled in quantitative methods, but also professionals skilled in composing and representing narratives about the lives of individual victims. A good example of a policy report that was turned into a powerful narrative is the *9/11 Commission Report* (National Commission on Terrorist Attacks Upon the United States, 2004), which was written by professional writers and published by a major publishing house, both of which contributed to its wide public consumption. Unique for a policy report, it was a best seller in 2004 and a finalist for the National Book Award.

On this last point, Paul Farmer (2005) wrote eloquently about the power of images, narratives, and first-person testimony to overcome our "failure of imagination" in contemplating the fate of distant, suffering people. Such documentation can, he asserted, render abstract struggles personal and help make human rights violations "real" to those unlikely to suffer them. But he is aware, as well, of the limitations of this information. He quoted Susan Sontag (2003), who cautioned that "as one can become habituated to harm in real life, one can become habituated to the harm of certain images" (p. 82). Sparking emotion with testimony and photographs, Farmer argued, is one thing; "linking them effectively, enduringly, to the broader project of promoting basic rights . . . is quite another" (p. 185). In short, he said, "serious social ills require in-depth analyses" (p. 185).

Further caveats about the use of atrocity images were expressed by Zelizer (1998), who argued that the recycling of images, such as photos of starving children in refugee camps, bears an eerie resemblance to photos from the Holocaust, which undermine their novelty and immediacy and can dull our responses. Similarly, Just (2008), reviewing the plethora of excellent books and movies on Darfur, observed that the horror they vividly depict should disgust us, but

> one effect of the extraordinary amount of knowledge we have about Darfur is that these stories eventually run together and lose their power to shock. . . . Repetition eventually numbs the moral imagination. . . . It is a terrible thing to admit, but the more information we consume about Darfur, the less shocking each piece of new information seems. . . . Ignorance is not the only ally of indifference; sometimes knowledge, too, blunts the heart and the will. (p. 41)

Another serious concern is the distributional effects of information conveyance that relies on images, narratives, and storytelling. The types of individuals and lifestyles that will trigger emotional connections may be implicitly affected by race, sexual orientation, gender, class, and the like. One must be especially concerned about a medium in which culturally disempowered groups often lose to more "compelling" stories of popular and culturally similar groups. Consider the case of Darfurees and the American news media. According to the Tyndall Report, which monitors American television coverage, ABC News allotted a total of 18 minutes on the Darfur genocide in its nightly newscasts in 2004, NBC had only 5 minutes, and CBS only 3 minutes. Martha Stewart received vastly greater coverage, as did Natalee Holloway, the American girl missing in Aruba.

VICTIM EMPOWERMENT

Another domain is victim empowerment. Where System 2 processes are systematically lacking, victims could be empowered to trigger a range of institutional responses such as initiating international court proceedings, placing an issue on the agenda of an international political body, or making a presentation as part of the deliberative process. Human rights organizations, including the UN Office of the High Commissioner for Human Rights, could personally involve victims in making such presentations or reading their organization's statement before such bodies. In the abstract, such measures risk biasing decision makers toward System 1 emotional responses, which would be inappropriate in certain decision-making forums. Regime designers would need to consider the conditions for crafting such interventions primarily to prod System 2 mechanisms into action when they are otherwise deficient.

Directly Promote System 2 Deliberation

Even when System 1's moral intuitions are distorted, human cognition can rely on the rational, deliberative mode of thinking characteristic of System 2. Where emotion and affect let us down, we still can be spurred into action if we can trigger a deliberative process capable of weighing the costs and benefits of possible intervention options. In short, institutional design should focus on ways to directly engage System 2 in the consideration of mass human rights violations.

The role of psychology in mediating our reactions to genocide may suggest the promise of a supplemental remedy, one that, paradoxically, is actually quite modest on its face—a "less is more" approach to the international legal regime combating genocide. Rather than solely focusing on obligations to *act*, international and domestic law should also require actors to *deliberate* and *reason* about actions to take in response

to genocide, thereby engaging System 2 cognition in order to overcome psychic numbing. The obligation to deliberate should apply to omissions (e.g., the failure to respond meaningfully to a genocide) as well as to acts. Psychological research indicates that this simple act of reasoned decision making may help overcome cognitive obstacles to intervention.

Can legal institutions in fact promote deliberation, either among policy makers or among the general public? Although the law is typically conceived as being concerned with action and not deliberation, institutional designers have taken just such an approach in a number of areas of law and policy, seeking to promote better outcomes not just by regulating the end result of the decision-making process but by regulating the process itself as well. One important example is the legal requirement in many countries that governmental agencies produce environmental impact statements before taking actions that might have deleterious environmental effects. These procedural requirements are often self-consciously deliberation-forcing mechanisms: they do not bar agency action that would harm the environment; they simply require that these effects be considered. And while the success of such laws in actually altering outcomes has been debated, advocates for the environment have at least taken them seriously enough to push for enforcement of such requirements in the courts, even without a guarantee that the ultimate policy decision will be affected.

A more broadly applicable example from U.S. administrative law is the requirement that cost-benefit analysis be performed in the course of deciding whether to regulate *or to not regulate*. While cost-benefit analysis was initially considered a means for achieving deregulatory results, recent developments in the administrative state have illustrated the analysis's potential for promoting the consideration of beneficial regulations (Hahn and Sunstein, 2002). Applied without a deregulatory bias, this policy might be viewed as a deliberation-forcing rule to insure that the government does not fail to consider potential welfare-promoting actions. Consider another example of deliberation-forcing devices in legislative affairs. In the landmark case *Doctors for Life International* (2006), the South African Constitutional Court enforced a constitutional provision requiring participatory democracy by ordering the legislature to hold public hearings and debates. The court drew inspiration from similar constitutional requirements in other countries. All of these examples demonstrate a concern with the quality of deliberation given to controversial government decisions and manifest an expectation that improved deliberation can result in improved decisions, even without mandating what the final decision itself must be.

These examples indicate that pursuing a deliberation-forcing approach to antigenocide efforts would not be unprecedented as a supplemental legal tool designed to overcome the cognitive obstacles in the way of interventions. Moreover, because it requires "only" deliberation, states may be more willing to take on such obligations. At the international level, an additional protocol to the Genocide Convention could compel states to respond to genocide by producing a detailed action plan, factoring in the likely costs and benefits of different types of intervention. At regular intervals, states could be required to justify their failure to act based on an updated assessment of costs and benefits. And the treaty could require high-visibility public presentation of these findings before both international and domestic audiences. The reporting requirements could also specify engagement at both the elite decision-making level (e.g., requiring the participation of the security establishment) and involvement at the popular level (e.g., requiring dissemination of information and hearings designed to reach the public). In addition, the UN Security Council could create a genocide committee to monitor and receive state reports and to ensure that state reports are timely and do not constitute foot-dragging. Such a committee would be analogous to the 1540 Committee, which was established to monitor and coordinate national nonproliferation efforts. Finally, at the national level, legislatures and executives can require hearings and reports evaluating the costs and benefits of intervention and nonintervention. The important point is that a procedural obligation to deliberate may be less onerous but more likely to yield meaningful substantive responses in the advent of genocide.

Conclusion

Drawing upon behavioral research and common observation, we argue here that we cannot depend only upon our moral intuitions to motivate us to take proper action against genocide and mass abuse of human rights. This analysis places the burden of response squarely upon moral argument and international law. The genocide convention was supposed to meet this need, but it has not been effective. It is time to reexamine this failure in light of the psychological deficiencies described here and design legal and institutional mechanisms that will compel us to respond to genocide and other mass harms with a degree of intensity that is commensurate with the high value we place on individual human lives.

The stakes are high. Failure to overcome psychic numbing may condemn us to witness another century of genocide and mass abuses of innocent people.

Notes

Portions of this chapter appeared earlier in the paper "If I Look at the Mass I Shall Never Act: Psychic Numbing and Genocide," which was published in *Judgment and Decision Making* (2007, *2*, 79–95). We wish to thank the William and Flora Hewlett Foundation and its president, Paul Brest, for support and encouragement in the research that has gone into this chapter. Additional support has been provided by the National Science Foundation through Grants SES-0649509 and SES-1024808. Many individuals have provided constructive criticisms and helpful suggestions on this work as well as other intellectual and logistical support. Among the many, Ellen Peters and Daniel Västfjäll deserve special thanks. Finally, this chapter has benefited greatly from the advice and comments of Dan Ariely and Cass Sunstein. David Zionts currently serves as Special Advisor to the Legal Adviser, U.S. Department of State. The views expressed here are his own and do not necessarily reflect those of the U.S. Department of State or the U.S. Government.

References

Associated Press. (1994, September 21). 38,000 shoes stand for loss in lethal year. *Register-Guard* (Eugene, OR), pp. 6A.

Baron, J. (1997). Confusion of relative and absolute risk in valuation. *Journal of Risk and Uncertainty*, *14*, 301–309.

Barrett, L. F., and Salovey, P. (Eds.) (2002). *The wisdom in feeling*. New York: Guildford.

Bartels, D. M., and Burnett, R. C. (2006). *Proportion dominance and mental representation: Construal of resources affects sensitivity to relative risk reduction*. Unpublished manuscript. Northwestern University, Evanston, IL.

Batson, C. D. (1990). How social an animal? The human capacity for caring. *American Psychologist*, *45*, 336–346.

Batson, C. D., O'Quin, K., Fultz, J., Vanderplas, M., and Isen, A. (1983). Self-reported distress and empathy and egoistic versus altruistic motivation for helping. *Journal of Personality and Social Psychology*, *45*, 706–718.

Clark, M. S., and Fiske, S. T. (Eds.) (1982). *Affect and cognition*. Hillsdale, NJ: Erlbaum.

Coke, J. S., Batson, C. D., and McDavis, K. (1978). Empathic mediation of helping: A two-stage model. *Journal of Personality and Social Psychology*, *36*, 752–766.

Dallaire, R. (2005). *Shake hands with the devil: The failure of humanity in Rwanda*. New York: Carrol and Graf.

Davis, K. E., Kingsbury, B., and Merry, S. E. (2010). *Indicators as a technology of global governance* (Report No. 2010/2). Retrieved from Institute for International Law and Justice website: http://www.iilj.org/publications/2010-2.Davis-Kingsbury-Merry.asp

Dehaene, S. (1997). *The number sense: How the mind creates mathematics*. New York: Oxford University Press.

Dickert, S., and Slovic, P. (2009). Attentional mechanisms in the generation of sympathy. *Judgment and Decision Making*, *4*, 297–306.

Dillard, A. (1999). *For the time being*. New York: Alfred A. Knopf.

Doctors for Life International v. Speaker of the National Assembly and Others, 12 BCLR 1399 (2006).

Eisenberg, N., and Miller, P. (1987). Empathy and prosocial behavior. *Psychological Bulletin*, *101*, 91–119.

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American Psychologist*, *49*(8), 709–724.

Farmer, P. (2005, March). Never again? Reflections on human values and human rights. *The Tanner Lectures on Human Values*. Salt Lake City: University of Utah. Retrieved from http://www.tannerlectures.utah.edu/lectures/documents/Farmer_2006.pdf

Fechner, G. T. (1912). Elements of psychophysics. *Classics in the history of psychology*. Retrieved from http://psychclassics.yorku.ca/Fechner/ (Original work published 1860)

Fetherstonhaugh, D., Slovic, P., Johnson, S. M., and Friedrich, J. (1997). Insensitivity to the value of human life: A study of psychophysical numbing. *Journal of Risk and Uncertainty*, *14*(3), 283–300.

Forgas, J. P. (Ed.) (2000). *Feeling and thinking: The role of affect in social cognition*. Cambridge: Cambridge University Press.

Friedrich, J., Barnes, P., Chapin, K., Dawson, I., Garst, V., and Kerr, D. (1999). Psychophysical numbing: When lives are valued less as the lives at risk increase. *Journal of Consumer Psychology*, *8*, 277–299.

Galanter, E. (1962). The direct measurement of utility and subjective probability. *American Journal of Psychology*, *75*, 208–220.

Glover, J. (2001). *Humanity: A moral history of the twentieth century*. New Haven, CT: Yale University Press.

Greene, J. D. (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *The neuroscience of morality: Emotion, brain disorders, and development* (Vol. 3, pp. 35–79). Cambridge, MA: MIT Press.

Hahn, R., and Sunstein, C. (2002). A new executive order for improving federal regulation? Deeper and wider cost-benefit analysis. *University of Pennsylvania Law Review*, *150*, 1489–1552.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834.

———. (2007). The new synthesis in moral psychology. *Science*, *316*, 998–1002.

Hamburg, D. A. (2008). *Preventing genocide: Practical*

steps toward early detection and effective action. Boulder, CO: Paradigm.

Hamilton, D. L., and Sherman, S. J. (1996). Perceiving persons and groups. *Psychological Review*, *103*(2), 336–355.

Jenni, K. E., and Loewenstein, G. (1997). Explaining the "identifiable victim effect." *Journal of Risk and Uncertainty*, *14*, 235–257.

Just, R. (2008, August 27). The truth will not set you free: Everything we know about Darfur, and everything we're not doing about it. *New Republic*, pp. 36–47.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*, 697–720.

Kahneman, D., and Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, and D. Kahneman (Eds.). *Heuristics and biases* (pp. 49–81). Cambridge: Cambridge University Press.

Kahneman, D., and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–291.

Kogut, T., and Ritov, I. (2005a). The "identified victim" effect: An identified group, or just a single individual? *Journal of Behavioral Decision Making*, *18*, 157–167.

———. (2005b). The singularity of identified victims in separate and joint evaluations. *Organizational Behavior and Human Decision Processes*, *97*, 106–116.

Kristof, N. D. (2007, May 10). Save the Darfur puppy. *New York Times*. Retrieved from http://select.nytimes.com/2007/05/10/opinion/10kr istof.html

Le Doux, J. (1996). *The emotional brain*. New York: Simon and Schuster.

Lifton, R. J. (1967). *Death in life: Survivors of Hiroshima*. New York: Random House.

Loewenstein, G., Weber, E. U., Hsee, C. K., and Welch, E. S. (2001). Risk as feelings. *Psychological Bulletin*, *127*, 267–286.

Mowrer, O. H. (1960). *Learning theory and behavior*. New York: John Wiley and Sons.

National Commission on Terrorist Attacks Upon the United States. (2004). *The 9/11 Commission Report*. Retrieved from http://www.9-11commission.gov/

Richter, E., and Stanton, G. (n.d.). *The precautionary principle: A brief for the Genocide Prevention Task Force*. Retrieved from http://www.genocidewatch.org/resources/bydrgreg orystanton.html

Schelling, T. C. (1968). The life you save may be your own. In S. B. Chase, Jr. (Ed.), *Problems in public expenditure analysis* (pp. 127–176). Washington, DC: Brookings Institution.

Schroeder, P., and Schroeder-Hildebrand, D. (2004). *Six million paper clips: The making of a children's holocaust museum*. Minneapolis: Kar-Ben Publishing.

Singer, P. (2007, March). Should we trust our moral intuitions? *Project Syndicate*. Retrieved from http://www.utilitarian.net/singer/by/200703--.htm

Slovic, P., Finucane, M. L., Peters, E., and MacGregor, D. G. (2002). The affect heuristic. In T. Gilovich, D. Griffin, and D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 397–420). New York: Cambridge University Press.

———. (2004). Risk as analysis and risk as feelings: Some thoughts about affect, reason, risk, and rationality. *Risk Analysis*, *24*, 1–12.

Slovic, S., and Slovic, P. (2004). Numbers and nerves: Toward an affective apprehension of environmental risk. *Whole Terrain*, *13*, 14–18.

Small, D. A., and Loewenstein, G. (2003). Helping *a* victim or helping *the* victim: Altruism and identifiability. *Journal of Risk and Uncertainty*, *26*, 5–16.

———. (2005). The devil you know: The effects of identifiability on punishment. *Journal of Behavioral Decision Making*, *18*, 311–318.

Small, D. A., Loewenstein, G., and Slovic, P. (2007). Sympathy and callousness: The impact of deliberative thought on donations to identifiable and statistical victims. *Organizational Behavior and Human Decision Processes*, *102*, 143–153.

Sontag, S. (2003). *Regarding the pain of others*. New York: Farrar, Straus and Giroux.

Stanovich, K. E., and West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*, 645–726.

Stevens, S. S. (1975). *Psychophysics*. New York: Wiley.

Susskind, J., Maurer, K., Thakkar, V., Hamilton, D. L. and Sherman, J. W. (1999). Perceiving individuals and groups: Expectancies, dispositional inferences, and causal attributions. *Journal of Personality and Social Psychology*, *76*(2), 181–191.

Tomkins, S. S. (1962). *Affect, imagery, and consciousness: Vol. 1. The positive affects*. New York: Springer.

———. (1963). *Affect, imagery, and consciousness: Vol. 2. The negative affects*. New York: Springer.

Totten, S. (Ed.). (2006). *Genocide in Darfur: Investigating the atrocities in the Sudan*. New York: Routledge.

Ubel, P. A., Baron, J., and Asch, D. A. (2001). Preference for equity as a framing effect. *Medical Decision Making*, *21*, 180–189.

Unger, P. (1996). *Living high and letting die: Our illusion of innocence*. New York: Oxford University Press.

U. S. Holocaust Memorial Museum. (2005). *Life after the Holocaust: Thomas Buergenthal—Personal history*. Retrieved from http://www.ushmm.org/wlc/en/media_oi.php?ModuleId=10007192&MediaId=5603

Van Berkum, J.J.A., Holleman, B., Nieuwland, M., Otten, M., and Murre, J. (2009). Right or wrong? The brain's fast response to morally objectionable statements. *Psychological Science*, *20*, 1092–1099. doi: 10.1111/j.1467-9280.2009.02411.x

Västfjäll, D., Peters, E., and Slovic, P. (2010). *Compassion fatigue: Donations and affect are greatest for a single child in need*. Manuscript in preparation.

Vedantam, S. (2010). *The hidden brain: How our unconscious minds elect presidents, control markets, wage wars, and save our lives*. New York: Spiegel and Grau.

von Lang, J. (Ed.). (1983). *Eichmann interrogated: Transcripts from the archives of the Israeli police*. New York: Farrar, Straus and Giroux.

Weber, E. H. (1834). *De pulsu, resorptione, auditu et tactu*. Leipzig: Koehler.

Wheeler, N. J. (2005). A victory for common humanity? The responsibility to protect and the 2005 World Summit. *Journal of International Law and International Relations*, *2*, 95–105.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist*, *35*, 151–175.

Zelizer, B. (1998). *Remembering to forget: Holocaust memory through the camera's eye*. Chicago: University of Chicago Press.

CHAPTER 8

Eyewitness Identification and the Legal System

NANCY K. STEBLAY

ELIZABETH F. LOFTUS

Anthony Capozzi was completely exonerated in 2007 from his earlier rape convictions, just one of nineteen wrongfully convicted persons exonerated of their crimes in that year alone through work of the Innocence Project. At the trial in 1987, the victims positively identified the Buffalo New York man as their attacker, and Capozzi, convicted of two rapes, spent the next twenty years in prison. Postconviction DNA testing of evidence that had been collected from the victims back in 1985 and saved in a hospital drawer proved that the real rapist, a man currently awaiting trial for murder, had committed the crimes for which Capozzi served time. On its website, the Innocence Project reports Capozzi's case and over 200 other wrongful convictions overturned by DNA evidence (http://www.innocenceproject.org/Content/488.php).

How were two-hundred-plus innocent persons like Capozzi erroneously convicted? False arrest and wrongful conviction can result from many types of errors. If there were no discernable patterns of risk factors for error, the justice system would probably be at a loss to offer solutions, perhaps judging these events as the unavoidable fallout of an imperfect justice system. However, this group of exonerations does present a clear pattern. Estimates place faulty eyewitness memory as having been involved in at least 75% of DNA exoneration cases, far more than faulty forensic evidence, bad informant testimony, false confessions, or any other cause (Garrett, 2008; Wells, Memon, and Penrod, 2006). Thus it behooves us to learn as much as we can about the science of eyewitness memory and to draw lessons for improving the justice system whenever eyewitnesses are playing a crucial role.

Research on eyewitness memory has been cited as having produced one of the most successful collaborations between psychological science and the legal system (Doyle, 2005), and a long-term contributor to the scientific literature recently proclaimed that "the science of eyewitness testimony has come of age" (Sporer, 2006, p. i). In this chapter, we will describe the nature and content of the massive research effort, the changes in legal policy spurred by the collaboration between those in science and those in the legal field, and the challenges ahead as memory research continues to be applied to public policy.

Eyewitness Memory Principles

Human memory, more generally, has been studied by psychologists for more than one hundred years, who have created a broad theoretical and empirical foundation for understanding the memory processes of interest to the legal system—in particular, eyewitness experience. Psychology's specific interest in the topic of eyewitness memory spans a hundred years as well. The writings of Hugo Munsterberg (1908) brought early attention to intriguing intersections between the young science of psychology and the established discipline of law (Doyle, 2005). The most recent thirty-five years have seen intense experimental examination of eyewitness issues and productive application of memory science to legal cases and policy. This activity has illuminated the task of psychological science to be one not only of establishing and disseminating new knowledge about memory function, but also one of correcting many memory myths held by both professionals and laypersons and articulating the role of psychological science in policy considerations. The audience for eyewitness research now includes lawyers, judges and juries, legislators, law enforcement, the media, and policy makers.

Five essential memory principles can provide a brief primer of eyewitness performance: *memory loss, memory construction, the misinformation effect, social*

influence, and *confidence inflation*. *Memory loss*, especially the process of forgetting, is perhaps the easiest concept to grasp, because we all recognize that the clarity of past life events fades with time. Instead, it is the complex process through which we attempt to *remember* persons and events that is often less understood by laypersons, a gap in knowledge that has presented challenges for memory scientists as they bring forward laboratory results to sometimes-resistant audiences. Scientists now recognize that memories do not lie hidden in pristine and easily recoverable form; rather, memory research has provided a much more sophisticated and nuanced understanding—that an experienced event is initially encoded incompletely into memory and later recalled through a *constructive memory process* that blends true recollection with intruding nonmemory features. An individual's beliefs, desires, and imagination can fuel misremembering, and information from external sources will exacerbate false recollection.

Research from the last three decades has established beyond question that information from outside of memory can be incorporated into a convincing "memory experience" (dubbed the *misinformation effect*) as memory gaps are unknowingly and effortlessly filled (Loftus, 2005). More dramatically, whole episodic memories can be constructed from imagination in the absence of true experience. Indeed, people have been led to believe that in childhood they were lost in a shopping mall for an extended time, broke a window and cut themselves, were attacked by an animal, or endured other experiences that would have been upsetting had they actually occurred (see, for example, Mazzoni, 2007). Thousands of experiments involving tens of thousands of participants provide documentation of the breadth and regularity of such disruptive memory effects. A leading memory researcher, Daniel Schacter, has fittingly referred to memory's "fragile power" (1996, 2001): the same remarkable brain capacity that allows for elaborate learning and effective social interaction and that holds each person's unique personal history—the essence of human identity—is also extremely vulnerable to misremembering and error.

Scientists recognize that eyewitness experience is not just a memory phenomenon, it also reflects social forces. Social and cognitive psychologists have established that memories can be enormously affected by even very subtle and unintentional verbal and nonverbal communications from other people. The process of *normative social influence* conveys expectations from one person to another about proper or modal behavior in a given situation (examples from a legal context include the eyewitness who feels pressure to make a choice from a police lineup or the patient who acquiesces to employ imagination during therapy). In addition, *informational social influence* provides seemingly useful knowledge to the recipient (e.g., that the suspect is in this lineup; that hypnosis will help recall) that subsequently affects perceptions, actions, and beliefs (Deutsch and Gerard, 1955). Appreciation of social influence principles is apparent in revised techniques for effective forensic interviews (e.g., Geiselman et al., 1985, 1986). More specific to eyewitness identification itself, current recommendations include an explicit "I don't know" response option for the eyewitness and a cautionary instruction that reminds the witness that the "offender you saw *may or may not* be in this lineup" (Steblay, 1997). This instruction has been shown to significantly reduce misidentifications, presumably by altering normative and social influences on the witness.

Interpersonal expectancy effects, the unintentional transfer of beliefs through social influence, occur across a broad set of human interactions (Harris and Rosenthal, 1985; Rosenthal, 2002; Rosenthal and Rubin, 1978). For example, within the science community, the expectations held by a researcher are recognized as threats to the integrity of research results. The well-known remedy is a double-blind method in which neither experimenter nor subject knows the subject's treatment condition. Double-blind studies are required in drug testing because we recognize that physician-researchers might inadvertently behave differently if they know a particular subject has been given the real drug rather than the placebo. Memory scientists similarly recommend appropriate methods for conducting forensic interviews such that interviewer knowledge is less likely to taint the direction and content of the questions pursued (e.g., Bruck and Ceci, 1997; Geiselman et al., 1985). The administration of double-blind lineups has also received substantial attention and is discussed below.

The powerful combination of postevent information delivered by a trustworthy nonblind authority can be observed in its remarkable impact on eyewitness *confidence*. An eyewitness who has received confirmatory feedback after her police lineup decision ("Good, you identified the suspect"), even if her choice was wrong, will show significantly more certainty about the identification and will report greater ease in making the identification than will a witness who did not receive feedback. In other words, confidence itself is highly malleable. Even more unsettling, the witness whose identification choice is "confirmed" will report distorted retrospective memory for subjective components of the crime event itself, claiming a better view and greater attention paid to the perpetrator (Douglass and Steblay, 2006; Wells and Bradfield, 1998; Wright and Skagerberg, 2007). The testimony of this

eyewitness is likely to be quite believable at trial because she truly accepts this version of reality. Investigators and jurors have been found to be strongly affected by confident, but sometimes inaccurate, witnesses (Bradfield and Wells, 2000; Brewer and Burke, 2002; Wells, Lindsay, and Ferguson, 1979). Confidence and accuracy are correlated; however, the relationship is easily corrupted.

In summary, five essentials of eyewitness memory—memory loss, memory construction, the misinformation effect, social influence, and confidence inflation—reveal the potential for memory to be contaminated and distorted and yet reported with great confidence. These lessons are immensely relevant to a legal system that depends on and believes in eyewitness veracity, and in which over 75,000 people become criminal defendants each year on the basis of eyewitness identifications (National Science Foundation, 1997). Although many applications of eyewitness memory for events are relevant to legal policy, we will focus on the illustrative example of eyewitness memory for faces and police lineup reform.

Key Events in the Growth of Lineup Reform

The Legal Environment

Eyewitness identification is persuasive evidence of criminal wrongdoing. Yet the courts recognize that an eyewitness may bring flawed recall to a police lineup and falsely incriminating evidence to court. In the 1960s, the United States Supreme Court began to institute safeguards to protect criminal defendants from misidentification and wrongful conviction. For example, in *United States v. Wade* (1967) the court held that the Sixth Amendment right to counsel applies to critical stages of pretrial proceedings, including the physical lineup procedure. The court recognized the "vagaries of eyewitness identification" and the "innumerable dangers and variable factors which might seriously, even crucially, derogate from a fair trial." The United States Supreme Court ruled in *Stovall v. Denno* (1967) that an unduly suggestive lineup constitutes a due process violation if it could lead to an irreparably mistaken identification. Therefore, a defendant could move to suppress identification testimony depending on the "totality of the circumstances" surrounding the testimony (p. 302). In *Simmons v. United States* (1968), the Court ruled that each potential due process violation during a lineup must be examined on the facts of the individual case. Lineups would be excluded from trial if the "procedure was so impermissibly suggestive as to give rise to a very substantial likelihood of irreparable misidentification" (p. 384).

Courtroom Testimony

For many years, psychological scientists have offered expert information to assist jurors and judges in assessing eyewitness accounts of criminal and civil events. A thorough explanation of the many factors that affect the accuracy of a witness's account of events or identification of a suspect can presumably be helpful to the triers of fact. Attorneys who have attempted to bring psychological testimony to trial have met with varying degrees of success, depending on the judge's determination as to the strength of the science underlying the proffered testimony, its perceived match to evidentiary standards, and particularly the need for the jury to hear the information. Wells et al. (2006) reported that prosecutors commonly use four core arguments against admission of expert testimony on eyewitness topics. The first, that the eyewitness literature is insufficient, today rarely prevails as a basis for excluding expert testimony. The three additional arguments are that such testimony invades the province of the jury to decipher the reliability of an eyewitness; that the research findings are simply a matter of common sense; and that the expert testimony is more prejudicial than probative, producing overly cautious jurors. Trial judges in most jurisdictions continue to use their discretionary powers to exclude expert testimony regarding the reliability of eyewitness identifications, often maintaining that scientific findings are not "beyond the ken" of the average juror (see Schmechel et al., 2006 for a summary of current case law on eyewitness research). In any individual case it may be difficult to calibrate the need for juror education on eyewitness topics and the extent to which expert testimony will appropriately remedy juror misconceptions. However, both laboratory research and surveys of eyewitness experts, judges, and prospective jurors suggest that, in general, jurors and judges relying on common-sense intuition often do not understand eyewitness memory processes and are likely to rely too heavily on eyewitnesses. Otherwise stated, eyewitness evidence without expert testimony is likely to exceed its probative value (Kassin et al., 2001; Schmechel et al., 2006; Wise and Safer, 2003). Wells and Hasel (2008) also make a persuasive argument that current police and court practice in itself is evidence that the justice system does not possess common knowledge of the psychological processes that affect the accuracy of eyewitness identification. For example, the legal system's continued trust in nonblind lineup administrators and in the ability of eyewitnesses to retrospectively assess how tainting variables (such as seeing the suspect on a news broadcast) affected their lineup decision illustrates that problems of eyewitness memory are not at all obvious to police and the court.

System and Estimator Variables

Expert trial testimony and the science that underlies it impart insight into the memory dynamics of eyewitnesses and provide probabilistic rules of likely outcomes for the typical witness in a given situation. However, after-the-fact determination of whether a particular eyewitness's experience is the rule or the exception is difficult, a problem inherent to the courtroom use of the eyewitness literature (Doyle, 2005). Responding at least in part to this circumstance, Wells (1978) outlined a framework for the consideration of eyewitness memory that became useful as both a theoretical and practical tool. Wells's insightful model of system and estimator variables helped to direct research attention to the possibilities for systemic change, highlighting the fact that the principles of human perception, memory, and social influence can illuminate not only the causes of faulty memory but also suggest preventive measures to preclude eyewitness failure.

DNA Exonerations

As noted earlier, new techniques of forensic DNA testing introduced in the 1990s and the formation of the Innocence Project in 1992 have helped to exonerate many wrongfully convicted individuals, to date more than two hundred (Innocence Project, 2006, 2007, 2008). Investigators, attorneys, and testifying witnesses who have helped to prosecute a later-exonerated individual realize with extreme regret that even well-intentioned by-the-book procedures can end very badly. Along with the horrific effects on the lives of violated innocent people and their loved ones, wrongful conviction leaves the true perpetrators on the streets to commit additional offenses. The reality of wrongful conviction also has the potential to erode public confidence in the justice system and citizens' sense of security. By the mid-1990s, law enforcement and the legal community could not help but look uneasily over their shoulders for past wrongful convictions.

The National Institute of Justice Guide

A decade ago, this confluence of events—eyewitness science, DNA exonerations, legal cases, and media coverage—propelled joint action among law enforcement, legal professionals, and eyewitness scientists. A group convened by Attorney General Janet Reno produced *Eyewitness Evidence: A Guide for Law Enforcement* (hereafter, "the guide"), which was published by the National Institute of Justice in 1999 (Technical Working Group for Eyewitness Accuracy, 1999; a training manual, *Eyewitness Evidence: A Trainer's Manual for Law Enforcement*, was published by the National Institute for Justice in 2003). Psychological science had shown that eyewitness reports are often unreliable and that unintentional police influence can affect witness lineup selections. The guide was a productive step toward remediation of this problem, providing science-based recommendations for effective collection of eyewitness evidence. Specific to police lineups, the guide offers clear advice: the eyewitness should be given unbiased lineup instruction ("The perpetrator may or may not be in this lineup"), lineups should be constructed fairly (e.g., foils matched to perpetrator description and the suspect not standing out in the lineup), and officers should record results in a prescribed manner. The guide did not endorse, but rather alerted law enforcement to, three developing refinements: sequential-lineup presentation format, double-blind lineup administration, and the use of computers for lineup delivery. In the years since, researchers have produced a solid body of laboratory evidence that supports the use of double-blind sequential lineups as a means to secure better-quality eyewitness evidence (Steblay and Dysart, 2008; Steblay et al., 2001), and most recently, computer delivery of photo lineups has been implemented by a small number of police departments.

The Reformed Lineup Protocol

Relative and Absolute Judgment

Standard police lineups present the eyewitness with all lineup members (e.g., six persons) at one time. Under this simultaneous format, eyewitnesses tend to compare lineup members to each other to determine which most closely resembles the offender in memory, a process of *relative judgment* (Wells, 1984). If the witness was able to encode a vivid memory of the perpetrator and this person is in the lineup (a *culprit-present* array), the likelihood of a positive and correct identification is increased. The concern, however, is whether the witness will recognize the absence of the offender when, in fact, the suspect is not the perpetrator. The DNA-exoneration cases—the majority of which were instances in which the actual offender was not in the lineup—illustrate exactly this problem: witness inability to correctly reject a *culprit-absent* lineup (Innocence Project, 2006). The results of controlled experiments predict a negative outcome when police unknowingly place an innocent suspect in a lineup. The witness may slip into the pursuit of which photo to choose, rather than a careful evaluation of whether the previously seen offender is one of the photos. Put

another way, the witness makes a relative judgment: "Number 5 is the closest compared to the others."

The impact of relative judgment when the offender is absent from the lineup was demonstrated convincingly by Wells (1993). Participant-witnesses to a staged crime were shown one of two versions of a lineup. When the perpetrator was present in a six-person lineup, 54% of the witnesses selected him. All witnesses had been given an unbiased cautionary instruction ("the perpetrator may or may not be in the lineup"), and 21% opted not to choose from the lineup. Now, the key question: What would happen when a second group of witnesses viewed the same lineup minus the perpetrator? If 54% of witnesses truly recognize the offender when he is present, this 54%—who would have identified the offender had he been in the lineup—should join the 21% who reject the lineup, producing a 75% "no-choice" response. What happened was quite different: only 32% of the witness responses landed in the "no-choice" category, these witnesses correctly rejecting the culprit-absent lineup. Sixty-eight percent of the witnesses chose from the lineup, most of the filler identifications falling on the photo that was the next-best match to the offender, placing this innocent suspect in jeopardy. Even in a culprit-absent lineup, it is likely that one lineup member will provide a better relative match to memory than the others, thereby drawing the attention of the eyewitness and increasing the risk of false identification.

Double-Blind Sequential Lineups

Most recently, scientists have advised police to use double-blind administration and a sequential photo presentation format for their lineup procedures (Wells et al., 2000). A meta-analytic review has demonstrated reliable laboratory outcomes with the use of a sequential procedure (Steblay and Dysart, 2008; Steblay et al., 2001). Witnesses who view a simultaneous lineup array are more likely to choose a photo from the lineup. When the perpetrator is present, this higher choosing rate may boost correct identifications, possibly aided by relative judgment. However, when the culprit is not in the lineup, an increased tendency to choose translates into greater risk of false identification. Recent cumulative data (Steblay and Dysart, 2008) show an average 8% fewer correct identifications of the culprit when the sequential is compared with the simultaneous format, but also an average 22% fewer identification errors. Thus, strategic use of a sequential versus simultaneous lineup format can be construed as a cost-benefit analysis. More precisely, the Bayesian likelihood ratio of a lineup procedure can be computed as the ratio of correct to mistaken identifications (Wells and Lindsay, 1980; Wells and Turtle, 1986). Wells (2006c) explains that the correct identification rate for a culprit-present condition can be divided by the average identification rate of any given person in the culprit-absent condition to produce a diagnosticity ratio; simply put, a ratio of hits to false alarms. The sequential lineup is more diagnostic of guilt (a ratio of 7.76) when the witness does make a choice than is the simultaneous lineup (ratio of 5.58). For police, the critical question is, Is the identification a good predictor of guilt? The blind-sequential-lineup procedure improves the odds that a suspect, if identified, is the actual culprit (Wells, 2006c) and thereby increases the probative value of the identification evidence (Lindsay et al., 2009).

Support for use of the double-blind component of the procedure is rooted in the broader psychological research about experimenter expectancy, which was discussed earlier. The exchange between the investigator and the eyewitness is ripe for potentially dangerous interpersonal influence. To help manage the risk for bias in identification procedures, eyewitness scientists recommend that lineup administrators be unaware—blind—to the identity of the suspect in the array. First noted as a lineup essential by Wells in 1988 and later reinforced by a broader group of scientists in lineup recommendations (Wells et al, 1998), there is wide agreement among eyewitness scientists that the administration of the double-blind lineup is crucial in eyewitness procedures (Douglass, Smith, and Fraser-Thill, 2005; Garrioch and Brimacombe, 2001; Haw and Fisher, 2004; McQuiston-Surrett, Malpass, and Tredoux, 2006; Phillips et al., 1999; Wells, 2006a; Wright and Skagerberg, 2007). A very recent experiment by Greathouse and Kovera (2009) explored the effect of the lineup administrator's knowledge of the suspect on the eyewitness's identification decisions, specifically with attention to the conditions under which administrator bias is likely to occur. In an eyewitness-identification paradigm in which the administrator's knowledge, lineup presentation format, and instruction bias were experimentally manipulated, the researchers found that administrator influence was significant under conditions that otherwise promote witness guessing. That is, witnesses were more likely to choose the suspect (apparently adopting a lower response criterion) when the lineup administrator knew the suspect, provided biased lineup instructions ("We have the suspect in custody, and would like to show you a photo lineup to see if you are able to identify him."), and presented photos simultaneously. When biasing factors were present to increase the likelihood of witness guessing, nonblind administrator behavior influenced the witness to choose the suspect.

Researchers use the phrase *double-blind sequential lineup* as shorthand for what is actually a collection of

rules that represent best practice for conducting eyewitness identifications. For example, the sequential procedure assumes a single-suspect model (only one suspect in the array) and that the lineup task is the first identification attempt by the witness. Furthermore, an effective sequential procedure includes the following features (see, e.g., Cutler and Penrod, 1988; Lindsay and Wells, 1985; Wells et al., 1998; Wells and Turtle, 1986):

A lineup consists of at least six members, five of whom are fillers unknown to the eyewitness, and all are chosen to match the witness's description of the perpetrator.

The suspect's position in the lineup is determined in a random manner.

An instruction to the witness advises that the perpetrator may or may not be in the collection of photos to be displayed (an "unbiased," or "cautionary," instruction).

The complete sequence of lineups is shown to the witness, and the witness is instructed that the complete series will be shown. Witness decision changes are recorded.

The witness is unaware of how many photos are in the sequence.

Photos are presented one at a time, with a decision made before examining the next.

The witness is not allowed to "go back" over the sequence or to place photos next to one another.

The officer displaying the photos does not know which photo depicts the suspect.

The witness is informed that the lineup administrator does not know which photo, if any, is the suspect.

An assessment of witness confidence is taken at the time of the identification and before feedback from police or others.

Lineup Reform

Scientific research has led to a cohesive lineup prototype that promises a significant improvement in eyewitness accuracy (Wells et al., 1998). The next step is to educate and to bring the recommendations into practice. Although courts in almost every jurisdiction have seen expert testimony about eyewitness identification over decades (Wells and Hasel, 2008), organized lineup reforms began to show up nationally just after 2000.

A powerful component of the lineup-reform effort has been the vivid and emotional testimony of the victims of wrongful conviction. Compelling presentations and writings by exonerees and crime victims have drawn national attention. Examples include Kirk Bloodsworth, the first death-row inmate to be exonerated by DNA evidence, who published a book and became a national spokesman for justice initiatives (Junkin, 1998). Similarly, Penny Beerntsen (2006) and Jennifer Thompson Cannino (2006), two victims and witnesses who unknowingly helped to convict the wrong men, are now educating audiences about eyewitness fallibility in the justice system. In addition, law enforcement officers, prosecutors, defense attorneys, and scientists, among many others, have become involved in the dissemination of information about eyewitness fallibility, wrongful conviction, and available remedies. Education underlies the reform effort, and many professionals are willing to provide the information to interested jurisdictions.

Below, we will briefly summarize the myriad ways in which lineup reform has occurred, in the hope that this may be instructive for other legal-reform efforts. Lineup reform has emerged from a number of catalysts and has been achieved through a variety of avenues: executive mandates, legislative actions, case law, and law enforcement initiatives, among them. The reforms to date illustrate a continuum of strategies, from jurisdictions in which detectives or police chiefs have initiated lineup reform ("bottom up") to those that change as a result of a mandate from high levels of government ("top down"). One early example of a grassroots orientation is provided by Lt. Ken Patenaude of the Northampton, Massachusetts Police Department, a long-time investigator and supervisor and a member of the National Institute of Justice Technical Working Group for Eyewitness Evidence. The changeover of Patenaude's own department to double-blind sequential lineups began at the ground level, when he developed and introduced a training program, providing long-needed structure and consistency in written procedures for securing eyewitness evidence (Patenaude, 2006). Police administrators monitored the implementation of the new sequential procedure for a year, at which time a survey of investigators revealed that they favored the new format. The department then changed its policy to mandate the sequential lineup format, at the same time noting a strong preference for double-blind lineup administration (Northampton Police Department, 2000). In 2003, the double-blind administration of photo arrays became mandatory as well, after concerns about

cost and personnel shortages failed to materialize. Patenaude's (2006) description of his department's transition to the new lineup procedures emphasizes the need to begin at the recruit level with proper and consistent training.[1]

Suffolk County (Boston, MA) followed a somewhat different route to lineup reform, spurred by discovery of wrongful convictions. A task force on eyewitness evidence was formed that brought together the perspectives of eyewitness scientists, law enforcement and administration, prosecutors, and defense attorneys. The task force issued its report with twenty-five recommendations to the Suffolk County District Attorney and the Boston Police Commissioner in 2004 (Suffolk County Task Force on Eyewitness Evidence, 2004); it was followed by reforms in lineup practice.

The combination of concerned, change-oriented law enforcement leaders, collaborative efforts, pilot testing, and effective police training has worked in a number of jurisdictions. In Minnesota, two counties independently developed year-long pilot programs. Under the direction of the county attorney, Amy Klobuchar, the Hennepin County Attorney's Office in Minneapolis tested the practicability of double-blind sequential lineups in four volunteer cities and became the first jurisdiction to collect data regarding eyewitness lineup decisions under double-blind sequential conditions and to document implementation issues. After one year, the changes were determined to be successful, and the new lineup protocol, along with a training DVD, was rolled out countywide (Klobuchar, Steblay, and Caligiuri, 2006). Next door, in Ramsey County (MN), County Attorney Susan Gaertner had also carefully examined the lineup literature and found the scientists' recommendations for double-blind sequential lineups sensible and potentially practicable. A pilot project was launched. Modifications were developed through experience, and all investigators found the new lineup procedures workable, as did the prosecutors who presented cases later in court (J. Schleh, personal communication, July 16, 2006). Ramsey County also developed written and DVD training materials as double-blind sequential procedures became the standard countywide. Among the benefits noted by Assistant County Attorney Jeanne Schleh were the increase in confidence in witness identifications, the reduced probability of misidentification, and the ability to insulate the prosecution from defense attack at trial since the new approach is consistent with best practices supported by established science (Schleh, 2006).

Local DNA exonerations can be the spur to action. New Jersey followed a highly prescriptive model in the wake of the *Cromedy* case—an eyewitness-evidence case in which after two trials, two convictions, and awaiting a third trial on appeal, a DNA test of biological evidence collected from the victim exonerated the defendant. Attorney General John Farmer turned to the lineup reforms recommended by researchers and approved new lineup procedures with safeguards exceeding those recommended by the National Institute of Justice (Doyle, 2005). Using the unique authority granted the attorney general in that state, Farmer implemented mandatory statewide guidelines, making New Jersey the first state to uniformly adopt double-blind sequential-lineup procedures (State of New Jersey, 2002).

At the state level, eyewitness-identification reform also has been attempted through a variety of legislative models (Ehlers, 2006). For example, a best-practices approach was used in Wisconsin, where the Training and Standards Bureau of the Wisconsin Department of Justice, working with the University of Wisconsin Law School, wrote model guidelines for law enforcement. Legislation passed in 2005 (State of Wisconsin Office of the Attorney General, 2005) and affirmed in 2006 required that each law enforcement agency adopt policies or guidelines (State of Wisconsin Office of the Attorney General, 2006).[2]

The formation of special state commissions has been used to learn about wrongful convictions and to identify remedies (see http://www.innocenceproject.org/fix/Eyewitness-Identification.php). A multistep process is typical. The first such group, the North Carolina Actual Innocence Commission, was established by the North Carolina Supreme Court in the aftermath of several high-profile DNA exonerations. The court decided that a permanent interdisciplinary study commission was needed, but one that was independent of the judiciary and had interdisciplinary participation of law enforcement, defense attorneys, social scientists, and judges (Garrett, 2006). The thirty-one-member commission created a series of recommendations in 2003 for state law-enforcement officers that left the details of implementation of these practices to the discretion of law enforcement. Later, state legislation mandated double-blind sequential procedures (Eyewitness ID Reform Act, 2007). To guide statewide efforts, model legislation is available through the Innocence Project.[3] The Justice Project (2007) also recently published a policy review and model guidelines.

One difficulty encountered with the legislative route is that precisely mandated reforms may need to be updated later as even better lineup revisions are developed (and this brings up the potentially messy revisitation of legislative actions). The

alternate route—leaving individual jurisdictions to find a solution—offers flexibility and local ownership but may result in protracted delays and less-than-effective outcomes. Even the middle ground—a task force and a pilot study—is not always successful. Wells (2006c) claims, based on his experience, that the actual costs of reform are minimal. However, he notes that the typical communication gap between eyewitness scientists and law enforcement, a tenacious police tradition, a lack of pressure from prosecutors and the court, and the disparate local control of law enforcement make lineup modifications difficult.

Resistance to Lineup Reform

Not all lineup-reform initiatives have gone smoothly. In 2002, Governor George Ryan's (Illinois) Commission on Capital Punishment, charged with ensuring the accuracy and justness of capital punishment, recommended the implementation of eyewitness-identification reforms (Governor's Commission on Capital Punishment, 2002). However, the proposed reforms were not popular with law enforcement (O'Toole, 2006). Resistance from police led to a compromise: a pilot program would be conducted by the Illinois State Police in which the new sequential lineup format would be compared to a simultaneous lineup format using "a protocol for the selection and administration of lineups which is practical, designed to elicit information for comparative evaluation purposes, and is consistent with objective scientific research methodology" (Capitol Punishment Reform Study Committee Act, 2003). The Illinois State Police ceded the pilot test to the Chicago Police Department, a group reportedly hostile to lineup reform (O'Toole, 2006) and to the direction by the general counsel for the Chicago police. Without rigorous scientific input as to the essentials of experimental design, three cities—Chicago, Joliet, and Evanston—collected data comparing a double-blind sequential-lineup protocol to the status quo (a relatively undefined nonblind simultaneous-lineup format). The 2006 report to the Illinois legislature (Mecklenburg, 2006) on the pilot program received substantial media attention, including the front page of the *New York Times* (Zernicke, 2006). Its surprising conclusion: the sequential double-blind lineup led to higher rates of false identification. The astute reader will appropriately note that dangerous false identifications of innocent suspects—the sort revealed by postconviction forensic DNA tests—cannot be ascertained in field lineup studies because the true guilt or innocence of the suspects is unknown. Mecklenburg's forceful use of the phrase "false identifications" (which, in fact, referred to nondangerous filler selections) and her decision to equate all suspect selections with true offender identifications served to inflame and confuse the subsequent discussion.

But the problems ran much deeper than semantics, and the Mecklenburg Report was critiqued and bitterly contested among scientists, lawyers, scholars, and policy makers (see e.g., Diamond, 2007; Doyle et al., 2006; Malpass, 2006; O'Toole, 2006; Sherman, 2006; Steblay, 2006; Sullivan, 2007; Wells, 2006a). Some scientists were quick to point out that the study's design was in numerous ways ineffectual by scientific standards and, above all, that the results were confounded by a fundamental design flaw; thus, the underlying reason for the obtained effects could not be determined (see Doyle et al., 2006; Steblay, 2006). More specifically, in one tested condition, the lineups were double-blind and sequential; in the other condition, the lineups were nonblind and simultaneous. Thus, it is unclear as to whether the outcomes were produced by the lineup format (sequential vs. simultaneous) or by administrator knowledge of the suspect (blind vs. nonblind); the variables were confounded. Furthermore, the specific results—higher suspect-identification rates and lower filler-selection rates in the nonblind simultaneous condition—are suggestive of administrator bias; critics maintain that it is not surprising to see that more witnesses chose suspects in a condition in which the lineup administrator knew who the suspect was. The conundrum is that the outcome data can be seen as evidence either of better lineup performance (and eyewitness accuracy) *or* of administrator influence and dangerous error introduced by the nonblind procedures. As noted earlier, there is no ground truth in the field (we do not know if the suspects are, in fact, perpetrators), thus the ambiguity of the results is increased. Diamond (2007) starkly states that this field test provides a classic example of what the law would deem not relevant; it provides exactly no probative evidence on the question at hand.

A remarkable step to attempt a resolution of the Illinois study controversy, and specifically to address the Mecklenburg Report that described its results, was quickly undertaken in 2006 by the Center for Modern Forensic Practice of John Jay College of Criminal Justice. This unprecedented action was succinctly explained by the center's director, James Doyle: "It's critical that criminal justice policy be based on sound science" (John Jay College of Criminal Justice, 2007). A panel of distinguished social scientists was convened to assess the Illinois field study, and they issued their report in February 2008 (Schacter et al., 2008). The panel of experts brought neutrality and outstanding collective expertise to the contentious issue.

The panel's clear determination was that the Illinois Eyewitness Identification Field Study was crippled by a design flaw that made the study's conclusions a dangerous basis for shaping public policy and the Mecklenburg Report unreliable in determining effective eyewitness-identification procedures. The Illinois study's fundamental design flaw "has devastating consequences for assessing the real-world implications of this particular study. . . . The design guaranteed that most outcomes would be difficult or impossible to interpret. The only way to sort this out is by conducting further studies" (Schacter et al., 2008, p. 4–5). Doyle summarized the panel's decision: "They found, unequivocally, that the Illinois report cannot be relied on to determine whether sequential double-blind procedures are effective. Most importantly, they recommend that future study of these procedures be designed in consultation with qualified scientists from the beginning, so that such studies can produce solid, reliable guidance for practitioners and policy makers" (Innocence Project, 2007; see also a series of 2008 articles by Cutler and Kovera; Mecklenburg, Bailey, and Larson; Ross and Malpass; Steblay; and Wells).

The author of the Mecklenburg Report stood by her initial conclusions (Mecklenburg, Bailey, and Larson, 2008a, 2008b). However, further suspicion of the Illinois data has been fueled by the refusal of the Chicago and Joliet Police Departments to share the underlying data of the report. A FOIA lawsuit was filed by the National Association of Criminal Defense Lawyers in conjunction with the MacArthur Justice Center at Northwestern University School of Law in Chicago (Jaksic, 2007). One objective of the lawsuit was to obtain previously unexamined information regarding the identification history of each witness and suspect, as well as data regarding the relationship between the suspect and witness. One of the three cities (Evanston) cooperated with the lawsuit, providing data from its 100 pilot-study lineups. The Evanston data revealed an additional crucial design flaw in the project (Steblay, 2009)—the failure of effective random assignment of lineups to the two tested conditions—and added to confusion about what exactly was measured in the Illinois pilot program. More precisely, the Evanston nonblind, simultaneous (status quo) condition included significantly more verification and confirmatory lineups. These are lineups in which the eyewitness simply verified the identity of a perpetrator known by the witness prior to the crime (e.g., a boyfriend or neighbor) or confirmed with a second identification a suspect selection that the same witness had already made from an earlier lineup); not surprisingly, these types of lineups produce high suspect-identification rates and very low filler-selection rates. In line with this, the suspect identification rate for nonblind simultaneous lineups was significantly inflated, by 17.7 percentage points, through inclusion of such lineups, compared to the sequential lineup, which was virtually unaffected (1% inflation from verification/confirmatory lineups). The failure to randomly assign the lineups to the two experimental conditions caused the status quo to look better. In the end, the assorted methodological shortcomings of the Illinois pilot program undermined the claims of the Mecklenburg Report.

The courtroom has seen setbacks for eyewitness science resulting at least in part from the Illinois study. Not only has the Mecklenburg Report been used to justify the status quo for lineup procedures, it also has been employed more broadly to challenge expert testimony on eyewitness topics such as cross-race identification and stress effects. The Public Defender's Service for the District of Columbia reports that the Illinois study is cited in nearly every government brief opposing expert testimony on eyewitness-identification issues and is heavily relied upon by prosecutors as "evidence" that status quo procedures are superior and that what has been heavily tested in controlled laboratory settings simply does not hold true in the field (B. Hiltzheimer, personal communication, July 13, 2007).

The results from the Illinois pilot study have proved inconsequential for some jurisdictions as they continue their reforms, but others view the fallout from the study as substantial. In Illinois, there has been no lineup reform to date, although the Capital Punishment Reform Study Committee (2007) reaffirmed its recommendation of blind-lineup administration. (A detailed summary of the Illinois capital punishment reform effort has been recently published by the co-chair of the governor's commission, Thomas Sullivan, 2007). To the dismay of reform advocates, the Illinois field study has been used to defeat legislation in several states that were otherwise moving toward sequential double-blind as a standard practice (B. Hiltzheimer, personal communication, July 13, 2007). In Rhode Island, for example, the Illinois report was used three years in a row to stop identification reform legislation (M. DiLauro, Office of the Rhode Island Public Defender, personal communication, July 18, 2007). The Mecklenburg Report was raised in New Mexico in connection with a successful police and prosecutor effort to squash a reform bill there (Rozas, 2007).

Science in Public Policy

Three decades of careful, peer-reviewed, and published research detail scientific knowledge regarding eyewitness memory. The body of eyewitness research

has matured over this period, precisely in the manner and to the standards of high-quality science: a steadily growing body of rigorous tests from independent labs has revealed reliable principles of eyewitness memory and behavior. This theoretically grounded body of literature is widely accepted in the science community (Kassin et al., 2001). Moreover, eyewitness science has made strong use of the quantitative review method of meta-analysis, a technique of research synthesis that particularly lends itself to the scrutiny and requirements of law. Meta-analysis allows scientists and the law to see beyond individual studies to overall patterns in the data, the "forest rather than the trees," and provides quantitative indices about reliability, effect sizes, and error rates (see Blumenthal, 2007, for a discussion of meta-analysis in legal policy). Peer-reviewed meta-analyses are particularly crucial as a means for employing the valuable self-correcting nature of scientific study. The eyewitness literature on a selection of both estimator and system variables has benefited from meta-analytic reviews.[4]

Given this "good news," the next question is Where does eyewitness science go from here?

Method Matters

The key role of eyewitness scientists is to keep good science front and center—in the laboratory, in the field, and in the public policy arena. As lineup reform moves forward, scientists must lead with their most valuable and productive attribute: adherence to sound scientific method and the logic of effective experimentation.

The future research agenda surely should, and will, involve field experiments, and method matters no less in the field than in the lab. Field studies bring unique strengths to research efforts, capturing eyewitness decisions not only in the most forensically relevant settings but also under circumstances that often lack the control and precision found in the laboratory. A primary and substantial challenge for eyewitness field tests is in the lack of available ground truth—that is, without follow-up tests of additional strong evidence, such as DNA, we cannot be certain that the police suspect is indeed the culprit. In the lab, of course, we have information about this crucial dependent measure. Public-policy makers can benefit when scientists discern the appropriate fit of lab and field results into the growing mosaic of knowledge about eyewitness memory.

Scientific attention to upcoming lineup field research is necessary in at least two specific ways: to define the proper method for lineup field tests and to bring scientific expertise to the interpretation of field results. There are multiple means to gain knowledge from the field. First, archival and descriptive studies offer a picture of how a lineup technique operates within a specific jurisdiction; they provide a starting point for discussion about the practicability and effectiveness for securing eyewitness memory in that locale. Examples are available in the work of Slater (1994), Tollestrup, Turtle, and Yuille (1994), Wright and McDaid (1996), Valentine and Heaton (1999), Behrman and Davey (2001), Valentine, Pickering, and Darling (2003), Behrman and Richards (2005), Klobuchar, Steblay, and Caligiuri (2006), and Wright and Skagerberg (2007). An important discovery from the Behrman studies is that traditional lineups conducted by police reveal that approximately 20% of witnesses pick an innocent foil from the lineup. These are, of course, known mistaken identifications that indicate unreliable witness memory. An important facet of these field investigations has been the opportunity for scientists to uncover variables overlooked in previous laboratory experiments. Of particular interest is any factor that would negate the viability of lab-based lineup recommendations. Thus far, there has been no sign of such a crippling factor.

The persuasive appeal of a good field study—and the potential for destructive misreading of field results from a poor one—can be enormous. Audiences may lock onto the phrase *field study* and quickly surmise that a report about real eyewitnesses to real crimes working with real police officers is a study to trust, not only in its description of *what* the eyewitnesses did but also in its conclusion about *why* these behaviors occurred. In essence, a field study may be used to automatically and uncritically eclipse "not-field" lab data. It is the role of scientists to counter, and in the best of worlds prevent through peer review, unfortunate leaps of logic—particularly when causal inferences in field or lab studies are not well grounded. Prudent evaluation of field research can be found in some descriptive studies that have attempted to examine the impact on eyewitness decisions of crime incident features, such as weapon presence. In these cases, the researchers were careful to point out the dangers of comparing pseudoexperimental conditions. For example, weapon presence or absence may be confounded with the type of crime (fraud vs. robbery) and therefore also with differential witness attention, the quality of the culprit description, and the delay prior to lineup (see, e.g., Steblay, 2007; Tollestrup, Turtle, and Yuille, 1994). The difficulty of interpreting study results following nonrandom assignment is illustrated by a London research team who compared the more controlled environment of a lineup suite (among other aspects, where volunteer foils are readily available for construction of higher-quality live lineups) to a standard police station setting (where the lineup members are "picked off the street" for each single lineup; Wright and McDaid, 1996). The

researchers noted that the lineups assigned to the suite differed in important ways from those assigned to ordinary police stations, such as in the time elapsed since the crime event, the race of suspect, and the violence of the crime.

Recently, a good deal of thought has been given to a more complex line of field research: experiments that can directly compare competing lineup strategies. In the fall of 2006, the Center for Modern Forensic Practice and the American Judicature Society brought together top eyewitness scientists and legal experts to map the methodological requirements for conducting future field experiments. The considerations included the means to creatively bring vital components of experimental design to eyewitness research in the field. Such features include double-blind testing, true random assignment to experimental conditions, clear operational protocols for stimulus materials and presentation, standardized instructions to participants (witnesses), and transparent documentation of the eyewitness-identification experience. The rationale for bringing standard experimental design components into field lineup investigations matches that for laboratory research: conclusions can be generated from the study more directly and with greater confidence if the appropriate controls are instituted across comparison groups and if the lineup task has been structured to minimize extraneous influences on the witnesses' decisions. Efforts are now underway to run sound field experiments in a number of cities nationwide.

Even with a properly designed and executed study, however, the interpretation of field reports and field experiments is tricky. There is no parallel in the field to the laboratory's culprit-present and culprit-absent lineups because the true status of the suspect as guilty or innocent is unknown. Field identification of a lineup member may be due to an eyewitness's true recognition of the offender *or* an erroneous choice of an innocent suspect. The worst-case scenario—when a witness's selection is incorrectly judged to be accurate—is illustrated by many DNA-exoneration cases. And, as we know from the laboratory, suspect-identification rates can be pushed up (and filler picks reduced) by undesirable practices that encourage witnesses to guess when their memory is poor or that bias the lineup structure toward the suspect (e.g., poor filler photos, a suspect with incriminating clothing, or a suggestive photo background). Left with measures that offer no absolute standard of goodness, lineup outcomes must be evaluated cautiously, within the context of the study design and the estimated gains or losses in witness-decision accuracy that are likely from the procedures employed. The interpretation of a field test demands a sophisticated understanding of memory principles, clarity about the underlying local street practice, and an appreciation of what field data can and cannot tell us. Law enforcement and researchers must together explore the implications of the data for future practice (see Steblay, 2007 for additional discussion).

The inherent ambiguity of eyewitness decisions in the field severely limits our ability to assess field outcomes with precision, and this impediment is likely to frustrate audiences who look for immediate definitive answers in field reports. Wise policy makers will continue to circle back to the laboratory for clarification of eyewitness phenomena. The primary objective of eyewitness research is better access to witness memory, and the benefit of laboratory lineup research is its methodical identification of factors that reduce or enhance eyewitness accuracy. As noted by Schacter et al. (2008) "no single study can produce a final blueprint for procedural reform." Just as in the lab, confident knowledge about field lineup performance will develop as evidence grows and patterns converge across jurisdictions and between the laboratory and the field.[5] As in all science, cumulative evidence carries more weight than any single study. Scientifically, the long view is much preferred to the short.

The laboratory will continue to feed the theoretical and empirical growth of the principles and applied knowledge of eyewitness memory. To help this process along, law enforcement officials are in a good position to identify gaps in that knowledge. Field studies already have prompted another iteration of lab inquiries in order to fine-tune the current recommended lineup protocol and to ascertain how adjustments in police lineup procedures that meet the convenience or practical needs of a local jurisdiction (including the current initiative to introduce laptop lineup administration) might compromise or enhance witness accuracy. Collaboration between the field and the laboratory has the potential to be very productive. A procedural anomaly or a creative idea brought forward by law enforcement may itself become the subject of experimentation and policy review, and perhaps get expediently built into the research design in both the field and lab to determine its impact on eyewitness decisions. For example, some jurisdictions prefer that eyewitnesses be allowed multiple viewings of the sequential lineup. Hennepin County (MN) permitted multiple "laps" in its pilot study (laps allowed only at witness request). In the lab, witnesses were offered the same option. The findings converged: the field data showed increasing filler selections (known errors) with lineup laps, and lab data echoed this pattern, establishing that misidentifications increased by 26% following repeated viewing of the lineup. (Klobuchar, Steblay, and Caligiuri, 2006; Steblay, 2007). As noted by Diamond (2007), there are significant benefits to learning "when well-documented field investigations are combined with laboratory backup" (p. 13).

Among these benefits is that laboratory researchers can use field research to inform their efforts to achieve desirable levels of authenticity and ecological validity in the laboratory.

Untidiness in Policy Development

The role of science in public policy has its limits. The squad room, the courtroom, and the legislative meeting room each have idiosyncratic perspectives and agendas not always in full synchrony with those of scientists. Consider, for example, some of the primary points of resistance to lineup reform, which are an assortment of political and logistical issues: "It's not broken." "It will cost too much." "It will slow our investigations and weaken our prosecutions." "It's soft on crime." "It favors the defense." "We are professionals and know best." There is also the practical concern that the courts will essentially punish reform efforts by opening the door to appeals of cases based on the traditional lineup if new sequential lineups are mandated (Taslitz, 2006).[6] Scientists can sometimes find creative empirical means to address such concerns, but for the most part, law enforcement, legal professionals, and policy makers must deliberate, test, and resolve these challenges. The justice system will accrue the long-term benefits of eyewitness reform if it can find immediate ways to boost the short-term value and the ease of reform implementation in today's street investigations and crime prosecution. For example, one promising means to satisfy logistical concerns is the use of laptop computers for lineup delivery. With the laptop method, fillers can be selected by a computer program, and the presentation of choices can be easily randomized and presented "blind" to the eyewitness. Helpful, unbiased instructions can be guaranteed. Computer cameras can even record the session.

Scientists and policy makers share a common conundrum: both must make decisions under conditions of uncertainty, and uncertainty spurs disagreement. As with most policy considerations, some discussants will voice apprehension about policy change in the absence of more complete information. This is true of recent debate about lineup reform. In the case of lineup reform, however, it should be pointed out that existing police procedures were not based on scientific memory principles or empirical evidence of effectiveness. The legal system has conducted little if any research on eyewitness memory and has no scientific theory of memory processes (Wells et al., 2006). Psychology's accumulating laboratory and field data evaluate the status quo practices as well as new procedures, and it may not be wise to presume inherent superiority in traditional practice.

Two examples of unknowns that have challenged lineup reform can be cited. First, lab research supports a *sequential superiority effect*—a sequential lineup display produces significant reductions in false identifications—but for undetermined reasons also reveals some loss in correct identifications compared to the simultaneous lineup format. Thus, from a policy perspective, there is not a simple solution, but rather an underlying balance to be achieved between avoiding erroneous identifications and securing accurate identifications of the guilty.

For some law enforcement, the average drop in correct identifications with a sequential format (one estimate is 8%; Steblay and Dysart, 2008) is read as a criterion shift that indiscriminately inhibits choosing in the sequential array and results in unacceptable nonidentifications of the truly guilty. On the other hand, a number of scientists speculate that the average difference in correct identifications between the two formats is at least in part accounted for by lucky guesses of witnesses with weak memories; the relative judgment in simultaneous arrays helps these guesses land on correct identifications when the offender is present (see Lindsay et al., 2009; Penrod, 2003; Steblay and Dysart, 2008; Steblay et al., 2001; Wells 2006b). According to these scientists, sequential presentation better captures true recognition. Sequential format does not make witnesses just hesitant to choose; rather, the witness becomes desirably cautious about choosing *just anyone* (see, e.g., Gronlund, 2004; Lindsay et al., 2009).

No perfect lineup procedure has yet been designed. Yet, the double-blind sequential lineup is viewed by many as promoting a higher quality of eyewitness evidence. If this is so, perhaps it should be the bright-line standard applied for courtroom eyewitness evidence. Is it also well suited for all stages of investigative police work? This is a practical issue for future examination. A policy decision to implement the blind-sequential reform rests on imperfect knowledge, and all said, also on political and philosophical justice issues: What level of risk to innocent suspects is tolerable in order to net more offenders? What is an acceptable basis for eyewitness-identification evidence? On balance, is the status quo—or the advocated reforms—justifiable? Such decisions also must involve a wise and wider view of police investigatory practice. For example, Lindsay et al. (2009) remind us that a failure to obtain a lineup identification does not preclude conviction; a case can be made against a suspect with other evidence.

A second, related, example is the unknown rate of target-absent lineups in the field. New lineup procedures are considered superior because of their demonstrated ability to reduce the risk of false identification

when the perpetrator is not in the lineup. Critics have argued that the need for lineup reform is undercut if offender-absent lineups (lineups with innocent suspects) are rare in field practice. However, true rates of target-absent lineups in the field, false identifications, and erroneous convictions are unknown, and perhaps ultimately unknowable. Perhaps more to the point—and strategically useful—is the recognition that there are countless circumstances under which police might unknowingly place an innocent suspect in a lineup and that the rate of target-absent lineups probably varies substantially across different jurisdictional, investigator practices, and stages of crime investigation (see Lindsay et al., 2009, and Wells, 2006c for in-depth discussions of this issue). Scientists can help law enforcement by continuing to determine a multitude of practice refinements to increase the probability that true offenders will be the focus of police investigation—at a lineup and at other points in an investigation. For their part, law enforcement professionals will need to determine whether to adopt reforms even as this knowledge continues to grow and with the awareness that their current investigative net is likely to snare an unknown number of innocent suspects.

Final Remarks

We have focused on the applicability of eyewitness science to reforms regarding lineups and other identification procedures. But the science also has bearing on cases in which eyewitnesses testify about matters beyond those involving the identification of perpetrators. In criminal cases, eyewitnesses testify about myriad matters. What was the color of the getaway car? Who started the fight, and were the defendant's actions a result of self-defense? Moreover, witnesses testify from memory about many matters that arise in civil cases, for example, the details of accidents, recollections of doctor-patient interactions in medical malpractice cases or of conversations in security fraud cases, and instances of claimed recovered memory, to name but a few. Are there reforms awaiting our consideration that would make the memory evidence more reliable and the verdicts in these types of cases more just? With creative scientific research, improved education for triers of fact, and constructive input from legal and policy communities, these are areas for future policy enhancement.

Notes

1. Other individual jurisdictions have also reformed their lineup procedures in the past decade even as their state practices have not changed. Examples include Virginia Beach, VA; Chaska, MN (Klobuchar and Knight, 2005); and Santa Clara County, CA, where Deputy District Attorney David Angel stated: "Some people have said that [these reforms] would reduce valid identifications, or they would be too expensive or too difficult to implement, but these problems have not come forward. . . . There is compliance; the training is not difficult; good IDs are made, and presumably they're more accurate" (Yeung, 2003).

2. Following a somewhat different approach, the Virginia General Assembly in 2004 instructed the Virginia State Crime Commission and Department of Criminal Justice Services to create guidelines for improving lineup procedures in the commonwealth and to develop training requirements for local jurisdictions (Ehlers, 2006). In 2005, the Crime Commission's recommendations were enacted, requiring that police departments have written lineup policies and procedures.

3. The Innocence Project website (http://www.innocenceproject.org/fix/Eyewitness-Identification.php) provides information on exoneration cases, reasons for wrongful convictions, connections to scientific work, model legislation, and a listing of reforms. This site includes materials and descriptions of the Northampton and Boston lineup reforms, among others.

4. Examples of estimator variable meta-analyses include cross-race identification (Meissner and Brigham, 2001b), eyewitness accuracy and confidence (Sporer et al., 1995), eyewitness stress (Deffenbacher et al., 2004), weapon focus (Steblay, 1992), exposure duration, retention interval, and disguises (Shapiro and Penrod, 1986), system-variable reviews on postidentification feedback (Douglass and Steblay, 2006), mugshot-exposure effects (Deffenbacher, Bornstein, and Penrod 2006), lineup instructions (Steblay, 1997), lineup format (Steblay et al., 2001; Steblay and Dysart, 2008), showups (Steblay et al., 2003), forensic hypnosis (Steblay and Bothwell, 1994), the cognitive interview (Kohnken et al., 1999), and verbal overshadowing (Meissner and Brigham, 2001a).

5. An ancillary line of hybrid lab-field research has developed around testing for fairness of real lineups. A *mock witness procedure* requires lab participants, who have not seen the crime and are armed only with the culprit description provided by the real witness, to identify the suspect from the lineup. This procedure is typically used to evaluate individual lineups suspected of biased structure. An emerging use of this method is to analyze a sample of lineups from a jurisdiction of interest. For example, in the Minnesota pilot of double-blind sequential lineups, a mock witness procedure confirmed fair lineup construction through a sample of field lineups (Steblay, 2007).

6. Discord is particularly common on the topic of the administration of double-blind lineups. With rare exception, eyewitness scientists see the double-blind procedure as an absolute necessity to maintain the integrity of the lineup

evidence. The *double-blind* method serves a dual function in lineup-reform research, providing the necessary method for objective comparison of competing lineup strategies (e.g., to test sequential versus simultaneous formats) and also shielding the eyewitness's decision and sense of certainty from the threat or suspicion of unintentional administrator influence. Proponents of this reform recognize that double-blind administration will increase the perceived and real integrity of the eyewitness evidence. While there is no need to assume bad behavior or intentionality on the part of the investigator (this is a protection against the very human phenomenon of unintentional communication), opponents see the drive for double-blind methodology as an insult to the integrity of detectives and their ability to handle witness interviews. Both sides to the argument cite professionalism as a reason for their position.

References

Beerntsen, P. (2006). Transcript of Penny Beerntsen's speech at Reforming Eyewitness Identification Symposium. *Cardozo Public Law, Policy, and Ethics Journal*, *4*(2), 239–249.

Behrman, B. W., and Davey, S. L. (2001). Eyewitness identification in actual criminal cases: An archival analysis. *Law and Human Behavior*, *25*, 475–491.

Behrman, B. W., and Richards, R. E. (2005). Suspect/foil identification in actual crimes and in the laboratory: A reality monitoring analysis. *Law and Human Behavior*, *29*, 279–301.

Blumenthal, J. A. (2007). Meta-analysis: A primer for legal scholars. *Temple Law Review*, *80*(1), 201–244.

Bradfield, A. L., and Wells, G. L. (2000). The perceived validity of eyewitness identification testimony: A test of the five Biggers criteria. *Law and Human Behavior*, *24*, 581–594.

Brewer, N., and Burke, A. (2002). Effects of testimonial inconsistencies and eyewitness confidence on mock juror judgments. *Law and Human Behavior*, *26*, 353–364.

Bruck, M., and Ceci, S. J. (1997). The suggestibility of young children. *Current Directions in Psychological Science*, *6*(3), 75–79.

Cannino, J. T. (2006). Transcript of Jennifer Thompson Cannino's speech at Reforming Eyewitness Identification Symposium. *Cardozo Public Law, Policy, and Ethics Journal*, *4*(2), 251–269.

Capital Punishment Reform Study Committee Act of 2003, Ill. Comp. Stat. Ann. 7255/107A-10 (West 2006).

Capital Punishment Reform Study Committee (2007). *Third annual report.* (Illinois). Retrieved from http://www.icjia.org/public/pdf/dpsrc/CPRSC%20Third%20Annual%20Report.pdf

Cutler, B. L., and Kovera, M. B. (2008). Introduction to commentaries on the Illinois pilot study of lineup reforms, *Law and Human Behavior*, *32*, 1–2.

Cutler, B. L., and Penrod, S. D. (1988). Improving the reliability of eyewitness identification: Lineup construction and presentation. *Journal of Applied Psychology*, *73*, 281–290.

Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993).

Deffenbacher, K. A., Bornstein, B. H., and Penrod, S. D. (2006). Mugshot exposure effects: Retroactive interference, mugshot commitment, source confusion, and unconscious transference. *Law and Human Behavior*, *30*, 287–307.

Deffenbacher, K. A., Bornstein, B. H., Penrod, S. D., and McCorty, E. K. (2004). A meta-analytic review of the effects of high stress on eyewitness memory. *Law and Human Behavior*, *28*, 687–706.

Deutsch, M., and Gerard, H. B. (1955). A study of normative and informational influence upon individual judgment. *Journal of Abnormal and Social Psychology*, *51*, 629–636.

Diamond, S. S. (2007). Psychological contributions to evaluating witness testimony. In E. Borgida and S. Fiske (Eds.) *Beyond common sense: Psychological science in the courtroom*. Malden, MA: Blackwell Press.

Douglass, A., Smith, C., and Fraser-Thill, R. (2005). A problem with double-blind photospread procedures: Photospread administrators use one eyewitness's confidence to influence the identification of another eyewitness. *Law and Human Behavior*, *29*, 543–562.

Douglass, A. B., and Steblay, N. M. (2006). Memory distortion and eyewitnesses: A meta-analysis of the post-identification feedback effect. *Applied Cognitive Psychology*, *20*, 859–869.

Doyle, J. (2005). *True witness: Cops, courts, science and the battle against misidentification*. New York: Palgrave Macmillan.

Doyle, J. M., Penrod, S., Kovera, M. B., and Dysart, J. (2006). The street, the lab, the courtroom, the meeting room. *Public Interest Law Reporter*, *11*, 13–46.

Ehlers, S. (2006). Eyewitness ID reform legislation: Past, present, and future. Talk presented at the Litigating Eyewitness Identification Cases Conference, Washington D.C.

Eyewitness ID Reform Act, S. 725, Gen. Assem., Sess. 2007, (N.C.) . Retrieved from http://www.ncleg.net/gascripts/BillLookUp/BillLookUp.pl?Session=2007&BillID=s725

Garrett, B. L. (2006). *Aggregation in criminal law.* Paper 43. University of Virginia Law School: Public Law and Legal Theory Working Paper Series.

———. (2008). Judging innocence. *Columbia Law Review*, *108*(1), 55–142.

Garrioch, L., and Brimacombe, C.A.E. (2001). Lineup administrators' expectations: Their impact on eyewitness confidence. *Law and Human Behavior*, *25*, 299–314.

Geiselman, R. E., Fisher, R. P., MacKinnon, D., and Holland, H. (1985). Eyewitness memory enhancement in

the police interview: Cognitive retrieval mnemonics versus hypnosis. *Journal of Applied Psychology*, *70*, 401–412.

———. (1986). Enhancement of eyewitness memory with the cognitive interview. *American Journal of Psychology*, *99*, 385–401.

Governor's Commission on Capitol Punishment. (2002). *Report of the Governor's Commission on Capital Punishment*. Springfield, IL: State of Illinois. Retrieved from http://www.idoc.state.il.us/ccp/ccp/reports/Commission_report/complete_report.pdf

Greathouse, S. M., and Kovera, M. B. (2009). Instruction bias and lineup presentation moderate the effects of administrator knowledge on eyewitness identification. *Law and Human Behavior*, *33*, 70–82.

Gronlund, S. D. (2004). Sequential lineups: Shift in criterion or decision strategy? *Journal of Applied Psychology*, *89*, 362–368.

Harris, M. J., and Rosenthal, R. (1985). Mediation of interpersonal expectancy effects: 31 meta-analyses. *Psychological Bulletin*, *97*, 363–386.

Haw, R. and Fisher, R. P. (2004). Effects of administrator-witness contact on eyewitness identification accuracy. *Journal of Applied Psychology*, *89*, 1106–1112.

Innocence Project. (2006). Mistaken I.D. Retrieved from http://www.innocenceproject.org/docs/Mistaken_ID_FactSheet.pdf

———. (2007, July 12). Social scientists say Illinois identification report is unreliable [Web blog post]. Retrieved from http://www.innocenceproject.org/Content/Social_scientists_say_Illinois_identification_report_is_unreliable.php

———. (2008). Eyewitness identification. Retrieved from http://www.innocenceproject.org/fix/Eyewitness-Identification.php

Jaksic, V. (2007, April 9). States look at reforming lineup methods. *National Law Journal* (online).

John Jay College of Criminal Justice (2007, July 7). "Blue-ribbon" panel of experts calls for more—and better—research of important law enforcement practice. Press release. Retrieved from http://johnjay.jjay.cuny.edu/info/calendar/pressRelease/pressReleaseDetails.asp?ID=93

Junkin, T. (1998). *Bloodsworth: The true story of the first death row inmate exonerated by DNA*. Chapel Hill: Algonquin.

Justice Project. (2007). Eyewitness identification: A policy review. Retrieved from http://www.psychology.iastate.edu/~glwells/The_Justice Project_Eyewitness_Identification_ A_Policy_Review.pdf

Kassin, S. M., Tubb, V. A., Hosch, H. M., and Memon, A. (2001). On the "general acceptance" of eyewitness testimony research: A new survey of the experts. *American Psychologist*, *56*, 405–416.

Klobuchar, A., and Knight, S. (2005, January 12). New lineup procedures can reduce eyewitness mistakes. *Minneapolis Star Tribune*, p. 11A.

Klobuchar, A., Steblay, N., and Caligiuri, H. L. (2006). Improving eyewitness identifications: Hennepin County's blind sequential lineup pilot project. *Cardozo Public Law, Policy, and Ethics Journal*, *4*(2), 381–413.

Kohnken, G., Milne, R., Memon, A., and Bull, R. (1999). A meta-analysis on the effects of the cognitive interview. *Psychology, Crime, and Law*, *5*, 3–27.

Lindsay, R.C.L., Mansour, J. K., Beaudry, J. L., Leach, A. M., and Bertrand, M. I. (2009). Sequential lineup presentation: Patterns and policy. *Legal and Criminological Psychology*, *14*, 13–24.

Lindsay, R.C.L., and Wells, G. L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential presentation. *Journal of Applied Psychology*, *70*, 556–561.

Loftus, E. F. (2005). A 30-year investigation of the malleability of memory. *Learning and Memory*, *12*, 361–366.

Malpass, R. S. (2006). Notes on the Illinois Pilot Program on Sequential Double-Blind Identification Procedures. *Public Interest Law Reporter*, *11*(2), p 5–8, 39–41, 47.

Manson v. Baithwaite, 432 U.S. 98, 114 (1977).

Mazzoni, G. (2007). Did you witness demonic possession? A response time analysis of the relationship between event plausibility and autobiographical beliefs. *Psychonomic Bulletin and Review*, *14*, 277–281.

McQuiston-Surrett, D., Malpass, R. S., and Tredoux, C. G. (2006). Sequential vs. simultaneous lineups: A review of methods, data, and theory. *Psychology, Public Policy, and Law*, *12*(2), 147–169.

Mecklenburg, S. H. (2006). *Report to the legislature of the State of Illinois: The Illinois pilot program on double-blind, sequential lineup procedures.* Retrieved from http://www.chicagopolice.org/IL%20Pilot%20on%20Eyewitness%20ID.pdf

Mecklenburg, S. H., Bailey, P. J., and Larson, M. R. (2008a, October). Eyewitness identification: What chiefs need to know now. *The Police Chief*. Retrieved from http://www.policechiefmagazine.org/magazine/index.cfm?fuseaction=display_arch&article_id=1636&issue_id=102008

———. (2008b). The Illinois Field Study: A significant contribution to understanding real world eyewitnesses. *Law and Human Behavior*, *32*, 22–27.

Meissner, C. A., and Brigham, J.C. (2001a). A meta-analysis of the verbal overshadowing effect in face identifications. *Applied Cognitive Psychology*, *15*, 603–616.

———. (2001b). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*, 3–35.

Morgan, C. A., Hazlett, G., Doran, A., Garrett S., Hoyt, G. Thomas, P., Baronoski, M., and Southwick, S. M. (2004). Accuracy of eyewitness memory for persons encountered during exposure to highly intense stress. *International Journal of Law and Psychiatry*, *27*, 265–279.

Munsterberg, H. (1908). *On the witness stand*. New York: Doubleday.

National Science Foundation (1997, January 3). *False identification: New research seeks to inoculate eyewitnesses against errors.* Press release. Retrieved from http://www.nsf.gov/pubs/stis1997/pr971/pr971.txt

Neil v. Biggers, 409 U.S. 188, 199 (1972).

Northampton Police Department (2000). Eyewitness Identification Procedure. *Administration and Operations Manual* (Chapter 0-408). Northampton, MA. Retrieved from http://www.innocenceproject.org/docs/Northampton_MA_ID_Protocols.pdf

O'Toole, T. P. (2006, August). What's the matter with Illinois? How an opportunity was squandered to conduct an important study on eyewitness identification procedures. *The Champion*, pp. 18–23.

O'Toole, T. P., and Shay, G. (2006). Manson v. Brathwaite revisited: Towards a new rule of decision for due process challenges to eyewitness identification procedures. *Valparaiso University Law Review*, *41*, 109–148.

Patenaude, K. (2006). Police identification procedures: A time for change. *Cardozo Public Law, Policy, and Ethics Journal*, *4*(2), 415–419.

Penrod, S. (2003, Spring). Eyewitness identification evidence: How well are eyewitnesses and police performing? *Criminal Justice Magazine*, pp. 36–47, 54.

Phillips, M., McAuliff, B. D., Kovera, M. B., and Cutler, B. L. (1999). Double-blind lineup administration as a safeguard against investigator bias. *Journal of Applied Psychology*, *84*, 940–951.

Rosenthal, R. (2002). Covert communication in classrooms, clinics, courtrooms, and cubicles. *American Psychologist*, *57*, 834–849.

Rosenthal, R., and Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences*, *3*, 377–386

Ross, S. J., and Malpass, R. S. (2008). Moving forward: Response to "Studying eyewitnesses in the field." *Law and Human Behavior*, *32*, 16–21.

Rozas, A. (2007, July 30). Best police lineup format not yet ID'd. *Chicago Tribune*. Retrieved from http://articles.chicagotribune.com/2007-07-30/news/0707290206_1_illinois-study-lineup-death-penalty-reforms

Schacter, D. L. (1996). *Searching for memory: The brain, the mind, and the past*. New York: Basic Books.

———. (2001). *The seven sins of memory: How the mind forgets and remembers.* Boston: Houghton Mifflin.

Schacter, D. L., Dawes, R., Jacoby, L. L., Kahneman, D., Lempert, R., Roediger, H. L., and Rosenthal, R. (2008). Studying eyewitness investigations in the field. *Law and Human Behavior*, *32*, 3–5.

Schleh, J. (2006, March 9). Sequential photo lineups using an independent administrator in Ramsey County. Memorandum. Ramsey County, MN.

Schmechel, R. S., O'Toole, T. P., Easterly, C. E., and Loftus, E. F. (2006). Beyond the ken? Testing jurors' understanding of eyewitness reliability evidence. *Jurimetrics*, *46*, 177–214.

Shapiro, P., and Penrod, S. (1986). A meta-analysis of facial identification studies. *Psychological Bulletin*, *100*, 139–156.

Sherman, L. W. (2006). To develop and test: The inventive difference between evaluation and experimentation. *Journal of Experimental Criminology*, *2*, 393–406.

Simmons v. United States, 390 U.S. 377, 384 (1968).

Slater, A. (1994). Identification parades: A scientific evaluation. London, UK: Police Research Group, Home Office.

Sporer, S. L. (2006). The science of eyewitness testimony has come of age. *Psychological Science in the Public Interest*, *7*(2), i–ii.

Sporer, S. L., Penrod, S., Read, D., and Cutler, B. L. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness identification. *Psychological Bulletin*, *118*, 315–327.

State of New Jersey, Office of the Attorney General (2002, April 18). Attorney General guidelines for preparing and conducting photo and live lineup identification procedures. Retrieved from http://www.state.nj.us/lps/dcj/agguide/photoid.pdf

State of Wisconsin, Office of the Attorney General. (2006). Response to Chicago report on eyewitness identification procedures. Retrieved from http://www.doj.state.wi.us/dles/tns/ILRptResponse.pdf

———. (2010). Eyewitness identification procedure recommendations. Model Policy and Procedure for Eyewitness Identification. Retrieved from http://www.doj.state.wi.us/dles/tns/eyewitnesspublic.pdf

State v. Ledbetter, S.C. 17307, 275 Conn. 534, 881A.2d 290 (2005). Retrieved from http://www.nlada.org/forensics/for_lib/Documents/1142893341.14/Ledbetter%20SCOTUS%20petition.pdf

Steblay, N. (1992). A meta-analytic review of the weapon-focus effect. *Law and Human Behavior*, *16*, 413–424.

———. (1997). Social influence in eyewitness recall: A meta-analytic review of lineup instruction effects. *Law and Human Behavior*, *21*, 283–297.

———. (2006). Observations on the Illinois data. Retrieved from http://web.augsburg.edu/~steblay/ObservationsOnTheIllinoisData.pdf

———. (2007). *Double-blind sequential police lineup procedures: Toward an integrated laboratory and field practice perspective.* Final report to the National Institute of Justice for grant no. 2004-IJ-CX-0044.

———. (2008). Commentary on "Studying eyewitness investigations in the field": A look forward. *Law and Human Behavior*, *32*, 11–15.

———. (2009). *It's more complicated than that: Lessons from the Evanston, Illinois field data.* Paper presented

at the American Psychology-Law Society conference, San Antonio, TX.

Steblay, N., and Bothwell, R. (1994). Evidence for hypnotically refreshed testimony: The view from the laboratory. *Law and Human Behavior*, *18*, 635–652.

Steblay, N., and Dysart, J. (2008). *Seventy tests of the sequential superiority effect: A meta-analysis.* Manuscript submitted for publication.

Steblay, N., Dysart, J., Fulero, S., and Lindsay, R.C.L. (2001). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: A meta-analytic comparison. *Law and Human Behavior*, *25*, 459–473.

———. (2003). Eyewitness accuracy rates in police show-ups and lineup presentations: A meta-analytic comparison. *Law and Human Behavior*, *27*, 523–540.

Stovall v. Denno, 388 U.S. 301–302 (1967).

Suffolk County Task Force on Eyewitness Evidence. (2004). *Report of the task force on eyewitness evidence.* Retrieved from http://www.innocenceproject.org/docs/Suffolk_eyewitness.pdf

Sullivan, T. P. (2007). Efforts to improve the Illinois capital punishment system: Worth the cost? *University of Richmond Law Review*, *41*, 935–969.

Taslitz, A. E. (2006). Eyewitness identification, democratic deliberation, and the politics of science. *Cardozo Public Law, Policy, and Ethics Journal*, *4*(2), 271–325.

Technical Working Group for Eyewitness Accuracy. (1999). *Eyewitness evidence: A guide for law enforcement.* Research Report. Washington, DC: U.S. Department of Justice.

Tollestrup, P. A., Turtle, J. W., and Yuille, J. C. (1994). Actual victims and witnesses to robbery and fraud: an archival analysis. In D. F. Ross, J. D. Read, and M. P. Toglia (Eds.), *Adult eyewitness testimony: Current trends and developments* (pp. 144–160). Cambridge: Cambridge University Press.

United States v. Ash, 413 U.S. 300, 321 (1973).

United States v. Wade, 388 U.S. 218, 228 (1967).

Valentine, T., and Heaton, P. (1999). An evaluation of the fairness of police lineups and video identifications. *Applied Cognitive Psychology*, 13, S59–S72.

Valentine, T., Pickering, A., and Darling, S. (2003). Characteristics of eyewitness identification that predict the outcome of real lineups. *Applied Cognitive Psychology*, *17*, 969–993.

Virginia State Crime Commission. (2005). *Mistaken eyewitness identification*. Report of the Virginia State Crime Commission to the Governor and the General Assembly of Virginia. H.R. Doc. No. 40.

Wells, G. L. (1978). Applied eyewitness testimony research: System variables and estimator variables. *Journal of Personality and Social Psychology, 36*, 1546–1557.

———. (1984). The psychology of lineup identifications. *Journal of Applied Social Psychology*, *36*, 1546–1557.

———. (1988). *Eyewitness identification: A system handbook*. Toronto: Carswell Legal Publications.

———. (1993). What do we know about eyewitness identification? *American Psychologist*, *48*, 553–571.

———. (2006a). Comments on the Illinois Report. March 29. Retrieved from http://www.psychology.iastate.edu/faculty/gwells/Illinois_Report.pdf

———. (2006b). Does the sequential lineup reduce accurate identifications in addition to reducing mistaken identifications? Retrieved from www.psychology.iastate.edu/faculty/gwells/SequentialNotesonlossofhits.htm

———. (2006c). Eyewitness identification: Systemic reforms. *Wisconsin Law Review*, *2*, 615–643.

———. (2006d). An important note on field studies of eyewitness identifications from lineups: Filler identifications are "conditional proxy measures." Retrieved from http://www.psychology.iastate.edu/faculty/gwells

———. (2008). Field experiments on eyewitness identification: Towards a better understanding of pitfalls and prospects. *Law and Human Behavior*, *32*, 6–10.

Wells, G. L., and Bradfield, A. L. (1998). "Good, you identified the suspect": Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology*, *83*, 360–376.

Wells, G. L., and Hasel, L. E. (2008). Eyewitness identification: Issues in common knowledge and generalization. In E. Borgida and S. Fiske (Eds.), *Beyond common sense: Psychological science in the court room* (pp. 157–176). Oxford, Blackwell Publishing.

Wells, G. L., and Lindsay, R.C.L. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin*, *88*, 776–784.

Wells, G. L., Lindsay, R.C.L., and Ferguson, T. J. (1979). Accuracy, confidence, and juror perceptions in eyewitness identification. *Journal of Applied Psychology*, *64*, 440–448.

Wells, G. L., Malpass, R. S., Lindsay, R.C.L., Fisher, R. P., Turtle, J. W., and Fulero, S. M. (2000). From the lab to the police station: A successful application of eyewitness research. *American Psychologist*, *55*, 581–598.

Wells, G. L., Memon, A., and Penrod, S. D. (2006). Eyewitness evidence: Improving its probative value. *Psychological Science in the Public Interest*, *7*(2), 45–75.

Wells, G. L., and Quinlivan, D. S. (2009). Suggestive eyewitness identification procedures and the Supreme Court's reliability test in light of eyewitness science: 30 years later. *Law and Human Behavior*, *33*. doi: 10.1007/s1097-008-9130-3

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., and Brimacombe, C.A.E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior*, *22*, 603–647.

Wells, G. L., and Turtle, J. (1986). Eyewitness identification: The importance of lineup models. *Psychological Bulletin*, *99*, 320–329.

Wise, R. A., and Safer, M. A. (2003). A survey of judges' knowledge and beliefs about eyewitness testimony. *Court Review*, *40*(1), 6.

Wright, D. B., and McDaid, A. T. (1996). Comparing system and estimator variables using data from real lineups. *Applied Cognitive Psychology*, *10*, 75–84.

Wright, D. B., and Skagerberg, E. M. (2007) Post-identification feedback affects real witnesses. *Psychological Science*, *18*, 172–178.

Yeung, B. (2003, October 29). Innocence arrested. SFWeekly.com. Retrieved from http://www.sfweekly.com/2003-10-29/news/innocence-arrested/ (Additional information available at http://www.innocenceproject.org/docs/Santa_Clara_eyewitness.pdf)

Zernicke, K. (2006, April 19). Study fuels debate over police lineups. *New York Times*, p. 1.

CHAPTER 9

False Convictions

PHOEBE ELLSWORTH

SAM GROSS

False convictions have received a lot of attention in recent years. Two-hundred and forty-one prisoners have been released after DNA testing has proved their innocence, and hundreds of others have been released without DNA evidence. We now know quite a bit more about false convictions than we did thirty years ago—but there is much more that we do not know, and may never know.

Background

False Convictions and Exonerations

Conceptually, convicting an innocent person is a misclassification, an error caused by the difficulty of evaluating uncertain evidence about a past event. Few misclassifications, however, are as troubling. A false conviction may destroy the life of the innocent defendant and deeply damage the lives of those close to him. He is punished as cruelly as the worst among us, by the state, in public. He is deprived of the life he once led and labeled a criminal, perhaps a vicious predator. He knows that he is innocent; he tells the truth to the authorities, but they ignore him. And in the process they usually make another mistake: they fail to pursue the real criminal.

Historically, the dominant reaction to this problem has been denial. Judge Learned Hand expressed this view memorably in 1923: "Our [criminal] procedure has always been haunted by the ghost of the innocent man convicted. It is an unreal dream" (*United States v. Garsson*, 1923). Judge Hand, of course, knew that innocent people are sometimes convicted; his claim was that it is so extremely rare that the risk should not affect public policy. We still hear echoes of that view, but they are increasingly unconvincing.

The fundamental problem with false convictions is that they are extraordinarily hard to detect. By definition, we do not know when a conviction is wrong, or we would not make the error in the first place: if we had a general test for innocence, we would use it at trial. The same ignorance that causes false convictions makes them exceedingly difficult to study. The only ones we know about are exonerations, those rare cases in which a convicted criminal defendant is able to prove his innocence after the fact.

A handful of such cases were known when Judge Hand wrote in 1923. Nine years later, Edwin Borchard published *Convicting the Innocent*, his classic collection of 65 exonerations dating back to the nineteenth century (Borchard, 1932). In the decades that followed several similar collections were released (Frank and Frank, 1957; Gardner, 1952; Gross, 1987; Radin, 1964), culminating in Radelet and Bedau's compilation of 417 cases of American defendants who had been convicted of homicide or of other capital crimes in the nineteenth and twentieth centuries (Bedau and Radelet, 1987; Radelet, Bedau, and Putnam, 1992).

In the meantime, the rate of exonerations increased sharply, first in the mid-1970s, when the death penalty came back into use in the United States after a judicial hiatus (*Furman v. Georgia*, 1972; *Gregg v. Georgia*, 1976), and then again in 1989 when the first DNA exonerations occurred. As a result, there have been hundreds of exonerations in the United States in the past few decades. They have changed our view of the nature of the problem of false conviction and have had a substantial impact on the criminal justice system.

We focus on these recent exonerations, which fall into four sets:

- In January 1989, David Vasquez became the first of 241 American defendants to date to be exonerated by DNA evidence (Conners et al., 1996; Innocence Project, 2009).[1] Almost all of these exonerations involve rape, although in some cases the defendant was also convicted of another crime, usually murder.

- Since 1973, 135 defendants who were sentenced to death for murder have been exonerated and released. DNA evidence played a substantial role in 17 of these death-row exonerations (Death Penalty Information Center, 2009).
- From 1989 through 2003, at least 135 American defendants who were convicted of felonies but not sentenced to death were exonerated without the benefit of DNA evidence. Unlike the DNA and death-row exonerations, there is no authoritative list of such cases. The vast majority were from convictions for murder (78%) or rape (12%) (Gross et al., 2005).
- In the past ten years, between 140 and 200 innocent defendants were released in mass exonerations when three major police scandals came to light: two in Texas, in 2002 and 2003, and one in Los Angeles, in 1999. In each of these sets of cases, police officers were caught systematically framing innocent defendants for possession of illegal drugs or weapons (Gross, 2008).[2]

There have been other exonerations since 1973, but these four groups include the great majority of those that have been described in publicly available systematic collections. It is a small set of observations, perhaps 650 to 700 exonerations across the whole country over a 35-year period. It is not much to go on, but it is a lot more information than we had in 1990.

Before we proceed to what we have learned from these several hundred exonerations, we should say a few words about what we *do not* know.

First, since there is no test for the actual innocence of convicted defendants, we rely on a proxy: the actions of government officials when claims of innocence are raised. As we use the term, *exoneration* is an official act—a pardon, a dismissal or an acquittal—declaring a defendant not guilty of a crime for which he or she had been convicted, because new evidence of innocence that was not presented at trial required reconsideration of the case (Gross et al., 2005).

Some exonerated defendants are no doubt guilty of the crimes for which they were convicted, in whole or in part, but the number is probably very small. It is extremely difficult to obtain this sort of relief after a criminal conviction in America, and it usually requires overwhelming evidence. On the other hand, it is clear that countless false convictions are never discovered. That is true for entire categories of cases, as we will see, and even among cases where exonerations do sometimes occur, they frequently depend on blind luck.[3]

Second, we know next to nothing about false convictions for any crimes except rape and murder. These two crimes—the most serious violent felonies—account for only 2% of felony convictions (and a much smaller proportion of all criminal convictions), but 95% of exonerations. The main reason is simple. Since almost all exonerations require large investments of scarce resources, they are only actively pursued in the most serious of cases. The 340 defendants who were exonerated and released from 1989 through 2003 spent, on average, more than 10 years in prison. Most had been sentenced to death or life imprisonment, and more than three-quarters to at least 25 years in prison (Gross et al., 2005). By comparison, 30% of all convicted felons in 2004 were not incarcerated at all, and the average term for those who were was just over 3 years (Durose and Langan, 2007).

The disproportionate attention to the most extreme cases explains the comparatively high number of exonerations among murder convictions, and especially death sentences. For rape, of course, the availability of DNA evidence has made exonerations much more accessible and common than for other serious violent felonies, for example, armed robbery. Even so, rape exonerations generally occur in the cases with the most severe sentences. Of 121 rape defendants exonerated from 1989 through 2003, over 30% were sentenced to life imprisonment, and the median sentence for the remainder was 30 years; for all defendants convicted of rape in 2000, 10% received probation, and the median sentence for the rest was 7 years (Gross, 2008).

What mistaken convictions have we left out? Of course we do not know, but we can make some educated guesses. For example, the number of wrongful convictions for robbery must be far greater than the few that have been discovered. Almost all wrongful convictions in rape cases involve eyewitness misidentifications, which are largely limited to cases in which the criminal is a stranger to the victim, but robberies by strangers outnumber rapes by strangers by a factor of 10 or more (Gross et al., 2005). In a study conducted before the advent of DNA testing, most of the comparatively few eyewitness misidentification cases that led to exonerations were robberies, not rapes (Gross, 1987). It stands to reason that false convictions for robbery still outnumber those for rape, but very few of them show up among the exonerations because there is no definitive evidence of innocence that is comparable to DNA.

Base rates suggest that most false convictions probably occur among the two overlapping groups that dominate *all* criminal convictions: (1) Comparatively light sentences, typically for comparatively minor charges. As we have seen, such cases are all but entirely missing from exonerations. (2) Guilty pleas. Over 95% of criminal convictions in America are based on guilty pleas, usually as a result of plea bargains—but only

about 6% of exonerations are of defendants who pled guilty, and they are more similar to other exonerations than to guilty pleas in general. The average sentence for 20 defendants who pled guilty and were later exonerated between 1989 and 2003 was 46 years in prison, which is not surprising given that all but one were charged with rape or murder and all faced the death penalty or life imprisonment (Gross et al., 2005).

Here again, we have scraps of relevant information, enough to disprove the common belief that innocent defendants virtually never plead guilty to crimes they did not commit (Hoffman, 2007). We know about false convictions for illegal possession of drugs and guns in the context of the *mass* exonerations that followed the discoveries of three systematic schemes by police officers to frame innocent defendants. Most of these defendants pled guilty in return for sentences far lighter than those that might warrant the cost and work that are usually required to have a chance at an individual exoneration (Gross, 2008). But how often do innocent defendants plead guilty in order to receive light sentences in other, more common contexts? And in what sorts of cases? We don't have a clue.

The Frequency of False Convictions

As recently as 2007, Justice Antonin Scalia wrote in a concurring opinion in the Supreme Court that American criminal convictions have an "error rate of .027 percent—or, to put it another way, a success rate of 99.973 percent" (*Kansas v. Marsh*, 2006). A highly comforting assessment, if true—but of course, it is absurd. The error was derived by taking the number of exonerations we know about—almost all of which occur in a tiny minority of murders and aggravated rapes—and dividing it by the total of all felony convictions, from drug possession and burglary to car theft and income-tax evasion. To actually estimate the proportion of erroneous convictions, we need a well-defined group of cases within which we can identify all mistaken convictions, or at least a substantial proportion of them. It is hard to imagine how that might be done for criminal conviction generally; however, it may be possible to do so, at least roughly, for the two types of crimes for which exonerations are comparatively common: rape and capital murder.[4]

For rape, there are some systematic data (not yet analyzed) on false convictions. In Virginia, the Department of Forensic Science has discovered hundreds of files on rape cases from the 1970s and 1980s with untested biological evidence that could be used to obtain DNA profiles of the rapists. A careful study of this DNA archive, or of similar sets of files elsewhere, could produce a good estimate of the rate of false convictions for rape in that jurisdiction for the decade or so before pretrial DNA testing became routine. So far, all we have are the results of a preliminary run in Virginia: 2 false convictions out of 22 cases, or 9% of that tiny sample (Liptak, 2008).

Capital murder is different. It stands out from other crimes not because of any special evidentiary advantage in determining whether convictions were in error, but because far more attention and resources are devoted to death-penalty cases, before and after conviction. As a result, death sentences, which represent less than one-tenth of 1% of prison sentences, accounted for about 22% of the exonerations from 1979 through 2003, a disproportion of more than 250 to 1 (Gross and O'Brien, 2008). This suggests that a substantial proportion of innocent defendants who are sentenced to death are ultimately exonerated, perhaps a majority. If so, the rate of capital exoneration can be used as a lower bound for the rate of false conviction among death sentences. Gross and O'Brien (2008) calculated that 2.3% of all death sentences in the United States from 1973 through 1989 ended in exoneration (86/3792), and Risinger (2007) estimated that 3.3% of the defendants sentenced to death for rape murders from 1982 through 1989 were exonerated by DNA evidence; but as the researchers note, even among death sentences, the true proportion of false convictions must be higher than the observed proportion of exonerations, perhaps considerably higher.[5]

Can we generalize from the false-conviction rate for death sentences? One might suppose that the error rate for other crimes is likely to be at least as high, considering that fewer resources are devoted to less serious cases. On the other hand, Gross (1998) argued that the error rate for murder in general, and capital murder in particular, is likely to be greater than for other felonies because the authorities are under enormous pressure to solve these heinous crimes. As a result they sometimes pursue weak cases that would otherwise be dropped, cut corners, or rely on questionable evidence. Unfortunately, there are no data on this point one way or the other. What we do know is that among the most serious criminal convictions of all—death sentences—miscarriages of justice are, at a minimum, an uncommon but regular occurrence, like death from diabetes (3.1% of all deaths in the United States) or Alzheimer's disease (2.8%) (Heron, 2007).

Causes and Predictors of False Convictions

Several evidentiary and procedural factors recur among exonerations: eyewitness misidentification, false confession, fraud and error on the part of forensic analysts, perjury by jailhouse informants and other witnesses

who testify in exchange for substantial favors, misconduct by police and prosecutors, and incompetent representation by criminal defense attorneys. All of these factors have been examined by social scientists and legal researchers, some extensively.

Eyewitness error is the most common cause of false convictions. It occurs in most known cases (Garrett, 2008; Gross et al., 2005), and it is the one most thoroughly researched. Many factors that can minimize the likelihood of eyewitness error are within the control of the police (system variables, as Wells called them [1978]): obtaining an immediate detailed description of the suspect from the witness; careful choice of lineup members; instructions that caution the witness that the true culprit may not be in the lineup; presentation of the lineup by a person who does not know who the actual suspect is; carefully recording the content and timing of all communications between the police and the witness; and scrupulous refusal to communicate any information about the suspect to the witness. Laboratory studies have demonstrated that all of these factors and others can affect the testimony of the witness and the chances of misidentification (cf. Steblay and Loftus, this volume). Case studies confirm that these are the most common causes of error in false convictions that have come to light (e.g., McGonigle and Emily, 2008).

Approximately 250 false confessions have been reported since the late 1980s (Leo, 2008), and Garrett (2008) reported that they occurred in 15% of the cases of prisoners exonerated by DNA evidence. A series of laboratory studies by Saul Kassin demonstrates that ordinary people can be induced to confess to wrongdoing much more easily than is commonly believed, that tactics often used in police interrogations (such as lying about incriminating evidence) can increase the likelihood of false confessions, and that trained police investigators are not very good at distinguishing true confessions from false ones (Kassin, 2005). There is strong evidence from actual cases that suspects who are young or mentally impaired are particularly vulnerable to suggestive police tactics that encourage false confessions (Leo, 2009). Although the empirical record on false confessions is less extensive than it is for eyewitness misidentification, we know a good deal about the kinds of tactics that elicit false confessions, (Kassin, 2008), and prohibiting these tactics would certainly reduce their frequency.

Forensic error (Garrett and Neufeld, 2009), perjury by informants (Warden, 2004), and prosecutorial (Armstrong and Possley, 1999) and ineffective defense work (Scheck, Neufeld, and Dwyer, 2003) are not so subject to controlled experimentation but have frequently been found in cases of actual false convictions. Some of these problems are caused by overtaxed resources and heavy caseloads and might be solved by spending more money. But not all. For example, forensic labs that are run by police departments are less likely to conduct unbiased analyses than fully independent labs no matter how well funded. And prosecutorial misconduct that leads to newsworthy convictions is unlikely to be punished.

There is no doubt that all these factors contribute to many, probably most, false convictions. Most innocent defendants who were misidentified, for example, would not have been convicted if no eyewitness had identified them. But information from exonerations alone is limited, even when it is reinforced by the results of controlled experimental studies. Experimental studies have identified factors that lead to evidentiary mistakes (misidentifications, false confessions), and these mistakes frequently occur in known false convictions (e.g., Scheck, Neufeld, and Dwyer, 2003). But experimental studies cannot tell us which mistakes are most important for false convictions because they do not measure false convictions. It appears, for example, that many—probably most—misidentifications (Gross, 1987) and false confessions (Drizin and Leo, 2004) do not lead to the conviction of innocent people. To really understand the significance of these factors we need to know more about the investigatory and adjudicative processes that produce false convictions.

First, we only know about the causes of those false convictions that we know about. As we have seen, that means that any generalizations we make are effectively limited to rape and murder cases that go to trial. For example, some defendants who cannot afford to post bail are offered the choice of taking plea bargains and going home on probation or insisting on their innocence and remaining in jail. That dilemma may be a major cause of false convictions for innocent defendants who plead guilty (see, e.g., PBS, 2004, the Erma Faye Stewart case), but we have no data with which to test that hypothesis. And the false convictions that are produced by this process may involve the same evidentiary and procedural factors we have discussed—or they may not: many of these cases are decided on slight evidence with little procedure.

Second, the occurrence of one of these causal elements is rarely a sufficient description of the process that led to a wrongful conviction. For example, when an innocent defendant falsely confesses after 20 hours of intensive interrogation, we must ask, Why did the police believe he was guilty and invest so much time in wringing a confession out of him? And why did they trust a confession obtained under these circumstances?

Third, while these factors are *causes* of false conviction, they are not *predictors*. For example, eyewitness misidentification appears to be the most common cause of wrongful rape convictions, occurring in nearly 90% of rape exonerations. But what does that

really tell us? With a handful of exceptions, all rape exonerations so far have occurred in cases in which there was no pretrial DNA testing. In these cases, the victim was expected to identify the defendant, unless it was physically impossible because it was dark or her face was covered. If she failed to do so, the case usually fizzled. In other words, before DNA evidence, an eyewitness identification was all but essential for a rape case to be prosecuted at all. If all rape convictions involve eyewitness identification, then all rape exonerations necessarily involve misidentification. But if we can only infer the misidentification on the basis of the exoneration, the misidentification could not have been used as a predictor of innocence.

What about police procedures that might cause an eyewitness to pick the wrong person? Experimental studies demonstrate that misidentifications can easily be caused by suggestive identification procedures: a police officer who knows which of the subjects in a lineup is the real suspect may intentionally or unintentionally make that person salient to the witness in subtle or obvious ways; or a witness may be called to the police station and shown a person in handcuffs who vaguely resembles that witness's description of the criminal; or a witness who repeatedly fails to identify the suspect's picture in different photographic lineups may eventually pick him because of a cumulative sense of familiarity (Steblay and Loftus, this volume).

But do suggestive identification procedures *predict* false convictions? That is not so clear. Suggestive tactics may be pervasive, whereas false convictions are rare. For all we know, suggestive tactics are used just as often in accurate identifications as in mistaken identifications. We know from experimental research that suggestive tactics increase the number of mistaken identifications, but suggestive identification techniques can also lead to true convictions. They may be as likely to provide the impetus that motivates an irresolute witness to declare an accurate choice as they are to produce an inaccurate one.

The same logic applies to other common evidentiary causes of false convictions. For example, as with misidentification, we know that a confession is false only after the fact, when other evidence has established the defendant's innocence. And as with suggestive identification procedures, prolonged and grueling interrogation—or controversial techniques, such as falsely telling the suspect that there is incriminating eyewitness or fingerprint evidence, or suggesting that the reason he has no memory of the crime is that he may have blacked out—might be as likely or more likely to elicit confessions from guilty suspects as from innocent ones.

To identify actual predictors of false conviction we need information about factors that can be observed in advance, before we know whether a conviction is true or false. And we need that information not only for exonerations but also for some comparable set of true convictions as well. For the most part, such data do not exist, but a few patterns are clear enough to be apparent from comparisons between data on exonerations and statistics on rape and murder convictions in general. (1) Innocent African American men are more likely to be falsely convicted of rape than innocent white men, especially if the victim is white, probably because white Americans are much more likely to mistake one African American stranger for another than to confuse members of their own race (Meissner and Brigham, 2001). (2) Innocent teenagers accused of murder are more likely to falsely confess than are innocent adults. (3) Minority juveniles are more likely than white juveniles to be falsely convicted of rape or murder (Gross et al., 2005).

For death sentences, it is possible to make direct comparisons between true and false convictions because the available records (while far from perfect) are much more complete than for other criminal convictions. Gross and O'Brien (2008) compared death-row exonerations to a sample of executed capital defendants, with the assumption that almost all of those who were executed were guilty. They found that false capital convictions are more likely (4) if the defendant had little or no prior criminal record, (5) if the defendant did not confess, and (6) if the police investigation took a long time.

Social and Institutional Context

Overview

The common image of a false conviction is derived from the murder and rape exonerations that we know about: after a difficult and troubled investigation, an innocent defendant is convicted at trial for a heinous crime of violence and sentenced to death or life in prison. There is every reason to believe that few false convictions bear any resemblance to this picture. Ninety-eight percent of felony convictions, and a larger proportion of all criminal convictions, are for lesser crimes, mostly property crimes, drug crimes, and assaults. Ninety-five percent of felony convictions are based on guilty pleas, usually after perfunctory investigations. In that mundane context, false convictions are not dramatic errors caused by recklessness or serious misconduct but rather are commonplace events: inconspicuous mistakes in routine criminal cases that never get anything close to the level of attention that sometimes leads to exonerations.

What is more, even the most disturbing false convictions may have ordinary histories (Lofquist, 2001). Consider the case of Antonio Beaver. In 1996 a white

woman was the victim of a carjacking in St. Louis (Innocence Project, 2009). She described the criminal as a black man wearing a baseball cap with a gap between his front teeth and helped the policed draw a composite sketch. Beaver was picked up a week later because he resembled the composite: he had chipped teeth. He was placed in a lineup with three other men, where he was one of two men in the lineup wearing a baseball cap and the only one with visible dental defects. He was picked by the victim, convicted at trial—even though his fingerprints did not match those on the rear view mirror of the victim's car—and sentenced to 18 years in prison. Beaver was exonerated by DNA in 2007, after serving more than 10 years, because the victim wounded the real criminal with a screwdriver and he bled on the car seat. The actual robber was identified by his DNA and fingerprints; he was serving time for other crimes.

We tend to think that causes should be proportional to their consequences (Ross and Nisbett, 1991), so when a terrible disaster strikes, we search for a cause as dramatic as the tragedy that followed. That instinct is often false. After the Challenger space shuttle exploded in 1986, the official investigation concluded that the immediate cause was a decision by NASA managers—under bureaucratic and budgetary pressure—to proceed with the launch and override warnings from engineers of a potentially catastrophic risk. But as Vaughan (1996) demonstrated, there was nothing unusual about the launch decision. The managers decided to carry on in the face of a known danger, with the concurrence of the engineers, as they had on many other occasions. They broke no rules and followed the established practices of an organization in which it was common to classify some risks as "acceptable." Similar patterns of routine behavior may cause most false convictions, big and small.

This sort of everyday behavior was probably behind Antonio Beaver's tragedy. The lineup was obviously biased, but casual and suggestive lineups are common, perhaps the rule. Most likely, they only infrequently lead to false convictions. In many, if not most, cases the police do have the right guy; if they do not, the witness may not pick the innocent suspect despite the suggestive procedure, or the real criminal may turn up with the victim's wallet in his pocket, or the false suspect may have an iron-clad alibi (e.g., he was in jail at the time of the crime). In Beaver's case, the police ignored physical evidence from the scene—fingerprints from an unidentified person and DNA that was not tested for a decade—but that, too, is commonplace and usually harmless. The upshot was a case that drew no attention: a black man who claimed to be innocent was convicted of aggravated robbery on the basis of a single cross-racial identification at an imperfect lineup. Most such defendants are guilty, and when they are not, we almost never find out. Beaver lucked out: the real robber bled on the car seat, the car was recovered, and a blood swab was collected and preserved.

We are not suggesting that nothing can be done about false convictions. Common practices can and often should be changed. But there are costs, and choosing the most effective reforms is not easy, especially when there is so little information about the underlying problem.

The Structure of Criminal Investigation and Adjudication

Criminal cases in America proceed through several stages.

IDENTIFYING THE CRIMINAL

The first task in any criminal investigation is to identify the criminal. This can take any amount of time or none at all. At one extreme, identification may be instantaneous (as when a killer reports a homicide and confesses) or it may precede the crime: in a sting, for example, the suspect is identified *before* the crime is committed. At the other end of the continuum, some criminals—like the notorious Zodiac Killer, who terrorized northern California in the late 1960s—are never identified. However long it takes, at this stage the authorities are still trying to answer the question, Who did it? The answer, whenever it comes, marks a fundamental shift in focus: from an investigation of the crime to the pursuit and prosecution of the suspect; from figuring out what happened to building a case against the person who they believe did it.

ARREST AND CHARGING

Once the criminal is identified he must be apprehended and arrested. This usually happens soon after identification, but occasionally a suspect may remain at large for a long time, or forever. Arrest triggers another set of changes. Typically, this is the point at which a prosecutor first learns about the crime. (In a minority of cases prosecutors are involved earlier, either because the crime is unusually conspicuous or because the arrest is the product of a proactive investigation rather than an after-the-fact response to a reported crime.) The prosecutor decides what charges to file, if any, and presents them in court, at which point the formal process of American criminal litigation begins. The case becomes a lawsuit with the prosecutor as plaintiff and the suspect as defendant. The defendant appears in court and hears the charges; he

may be released pending trial, or he may be detained, usually because he cannot afford to post bail; and he gets a lawyer to defend him, usually an appointed lawyer paid by the state. The adversarial structure is now complete.

PRETRIAL SORTING

The next stage of criminal proceedings is often called pretrial bargaining, but that is misleading. It suggests that trial is the expected mode of resolving a criminal case, which is false. For example, of defendants charged with felonies in 2002 in the 75 largest American counties, only 4% went to trial whereas 65% pled guilty, overwhelmingly to felonies (Cohen and Reeves, 2006). Overall, about 95% of all felony convictions in the United States are obtained by guilty pleas, usually as a result of plea bargaining between defense attorneys and prosecutors; in 2002 the proportion of guilty pleas for state-court felonies ranged from 68% of murder convictions to 98% of drug possession convictions (Durose and Langan, 2004). In some unknown proportion of these guilty pleas, the defendants are innocent.

Plea bargains are not the only cases that end before trial. Nearly a quarter of all felony cases are dismissed by prosecutors, usually because they do not have enough evidence to get convictions in court (Durose and Langan, 2003). Some of these dismissals (again, we do not know how many) happen to benefit innocent defendants. In other cases, the charges are dropped before trial because of affirmative evidence of innocence. Judging from two studies that focus on specific causes of false convictions, an innocent defendant who is arrested is more likely to be discovered and let go before trial than to be acquitted at trial or exonerated after conviction. Gross (1987) collected data on 60 misidentification cases in the United States from 1967 through 1983; in 35 cases, the charges were dismissed before trial, and in 25, the defendants were exonerated after conviction at trial; there were no acquittals. And Drizin and Leo (2004) reported on 125 suspects who falsely confessed to felonies (overwhelmingly to murder) between 1971 and 2002: 10 were arrested but never charged, 64 had their charges dismissed before trial, 7 were acquitted at trial, and 44 were exonerated after conviction.

TRIAL

Trials are uncommon among criminal cases in America but are heavily overrepresented among exonerations: they account for about 5% of felony convictions but 94% of the exonerations we know about, a disproportion of more than 350 to 1. Trials are more frequent for the crimes that account for the great majority of exonerations—murder (32%) and rape (16%) (Durose and Langan, 2004)—but those charges may be more likely to produce exonerations in part *because* they are more likely to go to trial. The common image of an American criminal trial includes a jury, but about 60% are conducted by judges sitting alone. Either way, 80%-90% of felony defendants who go to trial are convicted.

Trial, of course, is a highly formal and adversarial affair. It is a show run by lawyers, and in criminal cases the dominant lawyer is the prosecutor, the official who represents the state, decides whether to file charges and for what crime, makes the plea offer that usually determines whether a case goes to trial or ends in a plea bargain, and, if a case does go to trial, presents the evidence gathered by the police. A prosecutor is legally and ethically bound to "seek justice," and in particular to avoid convicting the innocent, but her main role at trial is more concrete. Like the defense attorney (who has no general obligation to the cause of justice), she is an advocate whose goal is to win. Both sides are expected to follow the rules of ethics and procedure, but within those forgiving limits, their job is to present evidence and argument and to undercut their opponents' evidence in whatever manner seems most likely to succeed.

REVIEW

After trial, a convicted defendant may appeal, but the review he will get is limited. The basic form of review, direct appeal, is generally restricted to claims that the lower court committed procedural error. New evidence may not be presented. The appellate court may only consider evidence that was presented at trial and may not reevaluate the factual accuracy of the judgment of the judge or jury. Its sole role is to decide whether there were procedural errors at trial that were serious enough to require trying the case over again.[6] Appellate courts reach that conclusion in only a small fraction of criminal appeals, perhaps 5%–7% (Davis, 1982; Scalia, 2001). Despite the formal rules, there is a wealth of anecdotal evidence that judges are more likely to reverse a criminal conviction on "procedural" grounds if they have doubts about the defendant's guilt (Davis, 1982; Mathieson and Gross, 2004), but the effect on defendants who actually are innocent, if any, may not be large. Garrett (2008) looked at a sample of 121 noncapital DNA exonerations that had produced written opinions on appeal at some earlier stage of review. He found a comparatively high reversal rate, 9%, but it was essentially the same as the reversal rate for a matched group of noncapital murder and rape appeals, 10%, and, whatever the comparison,

91% of these innocent defendants had lost their appeals.[7]

Almost all exonerations occur outside the structure of direct appeal. Appellate review is not designed to deal with new evidence (Davis, 1982), and in most cases, the exonerating facts are discovered only years after the appeals have run their course. At that point the defendant may file a petition for discretionary *extraordinary relief*, asking a court to reopen his case in light of the newly discovered evidence, or he may ask the prosecutor to join him in such a petition and then dismiss the charges, or he may apply to the governor for a pardon. All of these options require substantial resources that are rarely available, since criminal defendants, who are almost always poor, have no right to appointed counsel after their direct appeal.

Obtaining relief on a claim of factual innocence is very difficult. The structure of appellate review in our legal culture reflects a deep reluctance to reconsider trial-court verdicts even in the light of substantial new evidence of error, a bias that is often justified by reference to the high value we place on the finality of judgments. In many cases a posttrial investigation has so thoroughly undermined a criminal conviction that it is clear that the defendant would be acquitted at a new trial, but no court is willing to exercise its discretion to reexamine the original conviction (see, e.g., Wells and Leo, 2008, describing the notorious Norfolk Four cases).

Other systems of appellate review may be more forgiving. In civil-law countries on the European continent the search for factual accuracy is considered an ongoing process, from trial through appeal. New evidence may be considered on appeal, trial witnesses may be recalled to provide additional testimony, and the factual conclusions of the trial court may be reconsidered and revised (Damaska, 1986). We do not know whether this more open system of review is more successful at identifying miscarriages of justice at trial.[8]

Wrongful Convictions and the Adversary System

False accusations occur in all legal systems, and all legal systems require some means of discovering them and preventing them from leading to false convictions. From the time the police identify a person as the criminal and make an arrest, the American criminal justice system is adversarial. Judges have little power to direct the investigation, call witnesses, or ask for additional evidence if they feel that what the attorneys have presented is ambiguous or incomplete. There is no official comparable to the *juge d'instruction* in France, whose sole task is to find the truth by searching for both incriminating and exculpatory evidence. Instead, the prosecutor focuses on incriminating evidence, and the defense on exculpatory evidence.

Proponents of the adversary system argue that when each side has a vested interest in finding every scrap of evidence that favors its position, the sum of the evidence is greater than if a single person investigated the case (Fuller, 1961; Thibaut and Walker, 1975). If the case reaches trial, all of the evidence the judge or jury hears is presented by the two adversaries, the prosecutor and the defense attorney. The role of the defense attorney is relatively straightforward: to get the best possible outcome for the client. The prosecutor has a dual role: first to decide whether the evidence is sufficient to charge the suspect with the crime, and then to organize the information into a winning case. Some scholars have argued that the motivation to win the case may interfere with the motivation to find the truth (Givelber, 2001; Strier, 1996). There are no useful data on rates or discovery of false convictions in adversarial versus nonadversarial legal systems—doubtless both could be improved. But the adversary system is the one we use in the United States, and in this section, we will describe several of its psychological and structural features that may undermine the successful discovery of innocent defendants.

CONFIRMATION BIAS

When we read news stories about the exoneration of innocent people, we are often disturbed by the flimsiness of the evidence that got them convicted in the first place. A single eyewitness identifies a man, and the case proceeds to trial and conviction even though nine coworkers testify that he was on the job fifty miles away, and they would be unlikely to make a mistake since he was the only black person in the work group and "stood out like a raisin in a bowl of rice" (the Lenel Geter case, described in Gross, 1987). In another case, a boy whose mother had just been murdered was detained for more than 24 hours and grilled for 8 hours by interrogators who told him, falsely, that he had failed a lie-detector test and that the reason he had no memory of committing the crime was that he probably blacked out; he came to think it might be true and confessed (Connery, 1977, the Peter Reilly case). If the evidence in these cases looks so implausible to us, why did the prosecutors believe it?

In other cases, even after apparently incontrovertible evidence proves that the defendant could not have committed the crime (e.g., a time-coded videotape shows him somewhere else; *Schlup v. Dello*, 1995) or a DNA match shows that the perpetrator was someone else (Frisbie and Garrett, 2005, the case of Rolando Cruz and Alejandro Hernandez), police

and prosecutors continue to insist that the men they arrested and convicted are guilty. How does this happen?

At some point in every successful case, investigators must identify a prime suspect and form a theory of the case. When this happens, police and prosecutors begin to make a commitment to their theory, and they become subject to confirmation bias—the tendency to notice, believe, seek, and remember evidence consistent with their theory, while overlooking, doubting, forgetting, and reinterpreting evidence to the contrary (Findley and Scott, 2006; Nickerson, 1998). Confirmation bias is not deliberate misconduct, nor is it the conscious preparation of an argument designed to persuade a jury. It is a normal tendency to construe the world according to one's preconceptions, and it has been found in average citizens, students, doctors, accountants, and other professionals. In criminal investigations, it can lead the investigator to interpret ambiguous evidence as consistent with the prime suspect's guilt, to explain away evidence that points to someone else, and to concentrate on the suspect when searching for additional evidence. "The prime suspect becomes the only suspect" (Tavris and Aronson, 2007, p. 137). As the investigation proceeds from seeking information to building a case, it becomes possible to ignore increasingly powerful indications that the prime suspect is the wrong person.

In a series of experimental studies, O'Brien (2009) gave participants a lengthy police file and, after they had reviewed the first half of the material, asked half of them to write down the name of their prime suspect. The other participants were not asked to state a hypothesis. The second half of the file included several pieces of information that raised doubts about the guilt of the prime suspect, as well as information that was consistent with guilt. After reading the entire file, participants were given a chance to ask for additional information. Those who had named a suspect were more likely to ask for information focused on that suspect rather than other possible suspects and to interpret ambiguous or inconsistent evidence so as to make it compatible with the suspect's guilt.

Confirmation bias affects investigators even when their sole task is to discover the truth—doctors, scientists, and no doubt, *juges d'instruction*. But the task of the police and prosecutor in an adversary system is not so simple and creates contradictory demands that exacerbate this bias. As the case proceeds from initial investigation to trial, their task shifts from finding the truth to building a case against the defendant that will persuade a judge or a jury. A persuasive case requires a coherent story (Pennington and Hastie, 1992), one without loose ends, gaps, or inconsistencies. Thus inconsistencies may be explained away or considered too trifling to communicate to the defense attorney or the jury, loose ends may be tied up, and in some cases gaps may be filled. Confident that they have caught the criminal, the authorities may inadvertently exert pressure on an eyewitness who is reluctant to make an identification or on a lab technician who cannot quite reach a conclusion. In the case of a suspect who refuses to talk, this pressure may be more intentional.

O'Brien followed up her studies of confirmation bias with a study that examined the effects of this dual role. Some participants simply named a suspect, while others were put in the role of prosecutors and were told that they would later have to persuade people that their prime suspect was in fact the criminal. Knowing that they would have to convince others that they were right led to an even stronger tendency to focus exclusively on the prime suspect, to interpret ambiguous evidence as consistent with his guilt, and to explain away inconsistent evidence.

FALSE CLAIMS OF INNOCENCE

As we have said, we do not generally know whether a criminal defendant is guilty or innocent—with one important qualification. In nearly every case, the defendant knows the truth. This private knowledge explains the special status we accord to confession, which has been called the queen of evidence. It makes it possible for our system of criminal adjudication to run almost exclusively on guilty pleas. And it means that innocent defendants can identify themselves to the authorities, and they do—all the time. Unfortunately, many guilty defendants also say they are innocent. Since we have strong reason to believe that the great majority of criminal defendants are guilty, true claims of innocence get lost in the crowd.

It is difficult to separate true claims of innocence from false ones in any context, but some features of the adversarial system make it worse. Once defense attorneys enter the picture they stop their clients from confessing—or from talking to the authorities at all; they take over all communication with the state. In that role they are expected to present their clients as innocent, if at all possible. But everybody who works in the system—prosecutors, police officers and judges—knows that this is playacting, that defense attorneys rarely believe their clients are innocent. Their job is to obtain the best outcomes for their clients, acquittal or dismissal if possible, even if the clients are guilty, and they usually are. Defense attorneys who succeed in saving "obviously guilty" clients from conviction are considered stars by their colleagues.

But what if the defendant really *is* innocent? The defense attorney, faced with dozens of spurious claims

of innocence, may not be able to detect the few that are true and rarely has the resources to conduct the sort of investigation necessary to provide convincing evidence. So defense attorneys frequently see their job as getting the best deal they can for the defendant without worrying too much about actual innocence.

PREPARATION FOR ADVERSARIAL TRIALS

We face a similar problem when it comes to presenting evidence at trial. We require witnesses to testify in public, in the presence of the defendant, following strange rules of procedure. To perform this tricky and unfamiliar role, a witness requires guidance, preparation by the lawyer who calls her. Such prepping is particularly important because her testimony includes cross-examination by an opposing lawyer whose job is to discredit her, whether or not she is telling the truth. Even truthful witnesses must be taught how to look and sound truthful; that is one of a trial attorney's most important tasks.

Adversarial preparation may produce coherent and convincing testimony, but it can also undercut accurate evaluation of the evidence at trial. A vague or uncertain witness is less persuasive than one who answers all questions without hesitation (Wells, Lindsay, and Ferguson, 1979); therefore, testimony is rehearsed and confidence is bolstered, sometimes beyond what is warranted. This process is particularly dangerous when it begins in the early stages of the investigation. The prosecutor and the police officers who work with an eyewitness are expected to help the witness identify the defendant in court with conviction and clarity. It seems in keeping with that role for an officer to tell a witness who has just tentatively picked the suspect from a lineup—"Congratulations, you got him!"—but the end result may be a misleadingly confident identification in court six months later (Wells and Bradfield, 1998).

So far what we have described is permissible witness preparation, as our system runs. But if your role as a police detective includes helping an eyewitness testify effectively, why not help her identify the defendant in the first place? It is a short step from shaping the identification *testimony* that a witness will give in court to helping that witness *make the identification* in a precinct station by steering her toward the defendant, especially if the detective has no doubt that the defendant is guilty but worries that the witness may ruin the case by failing to say so.

The same logic applies to other police procedures, such as interrogation, gathering information from snitches, and interpreting forensic evidence. If the police or prosecutors believe that they already know who the criminal is, the purpose of these procedures is not to find anything out but instead to produce evidence that will convince a judge and jury. Reforms designed to protect the innocent will seem misguided to law enforcement officials who use these procedures not to discover the criminal but to build a case that will convict him. If they see the reforms as obstacles to convicting the guilty, they are likely to resist them or try to circumvent their effects.

GENERATING FALSE NEGATIVES

A false positive is the inclusion of an object in a category where it does not belong: diagnosing a healthy person as depressed or diabetic, for example. A false negative is the exclusion of an object from a category where it does belong: diagnosing a depressed or diabetic person as healthy. In any classification system there is a tradeoff between false positives and false negatives. Procedures that reduce one type of error often increase the other. If there are twelve major symptoms of depression—insomnia, loss of interest, suicidal tendencies, and so on—a doctor who diagnoses a patient as depressed if she shows any one of the symptoms will mistakenly include many people who are not depressed: there will be too many false positives. A doctor who requires that a patient exhibit all twelve symptoms before prescribing treatment will mistakenly miss many people who are seriously depressed: there will be too many false negatives.

Those who seek to reduce wrongful convictions—false positives—must recognize that the same reforms might also reduce the number of convictions of suspects who are actually guilty. Misleading a suspect into believing that he has been identified by an eyewitness may cause an innocent person to make a false confession, but at least as often it may cause a guilty person to give up and confess the truth, thereby increasing the probability of an accurate conviction. Many of the proposed reforms may make all convictions more difficult to accomplish, not just convictions of innocent people.

Some innovations increase the identification of the innocent without diminishing the identification of the guilty—scientifically conducted DNA analysis is the shining example—but for most there is likely to be a tradeoff. Not even the excellent safeguards against suggestive lineup procedures proposed by the American Psychology-Law Society (Wells et al., 1998) are immune from this problem. These recommendations include blind lineups, informing the witness that the culprit might not be there, and fairly constructed lineups. But they could cause a hesitant but accurate witness to fail to identify a suspect, even though the same witness might have made the identification if suggestive procedures had been employed.

For a few reforms, such as sequential lineups (Wells, 2006), preliminary evidence indicates that the likelihood of increasing false negatives is small, but so far there is little research.

There are many policy reasons to forbid suggestive identification practices, but we cannot assume that an unbiased procedure always leads to the right result. If the police actually do know who committed the crime and can get a witness to identify the person, the resulting conviction is a true conviction. Videotaping interrogations and lineups is also an excellent idea, but not foolproof: an aggressive defense attorney may find pieces of the tape that would shake the jury's confidence in the result, whether or not that result is accurate. These reforms are important, and we endorse them, but they are not cost free.

The adversary system exacerbates this problem. Good defense lawyers will exploit any weaknesses or irregularities in the prosecution to cast doubt on the guilt of the truly guilty: their job is to generate false negatives, as the prosecutors well know. Witnesses shown a blind, unbiased lineup may be less confident than witnesses shown a biased lineup, may express uncertainty, or may not identify anyone at all. The defense attorney will make the most of these weaknesses, emphasizing the witness's failure to make a confident identification. The same is true for other reforms designed to minimize false convictions: the defense will use them to cast doubt on the guilt of all defendants. Most police and prosecutors prefer to keep their investigations confidential and resist reform efforts because they may provide ammunition to the defense. An adversary system is a contest, and the search for truth is often eclipsed by the desire to win.

Policy Implications

Basic Issues

The most dramatic development in the provision of intensive medical care in the past ten years is probably the use of checklists. The best known is a simple form that requires doctors to note that they have taken several time-honored steps to prevent infections when inserting bloodstream catheters: wash hands, clean patient's skin with disinfectant, cover patient with sterile drapes, and so forth. In a pilot project in Michigan hospitals in 2004 and 2005, the use of this checklist decreased the rate of infection by 66% over 3 months; in 18 months it saved $75 million and more than 1,500 lives (Pronovost et al., 2006). It seems that the best way to prevent bloodstream infections in intensive care units is not a new drug or better equipment but a procedure that greatly increases the odds that doctors and nurses will do what they are already supposed to do.

Almost every reform we suggest is some version of trying to get police, prosecutors, and defense attorneys to do what they are already supposed to do. But doing that effectively is far more difficult for false convictions than for infections. For one thing, we are crippled by our ignorance. We know that checklists reduce deaths in hospitals because we can observe that outcome directly and compare mortality rates across different treatment regimes, but (by definition) we never recognize false convictions when they occur, and we only occasionally discover them later on. For example, we have no idea how many innocent defendants plead guilty or which ones do so and under what circumstances, so we are unlikely to identify the variables that matter. And we cannot learn much from field experiments. We might test a plausible technique for reducing false guilty pleas, but since we still will not be able to tell which defendants are guilty and innocent, we will not know whether it works.

The fundamental reason for our pervasive ignorance is that guilt is a classification based on imperfect information. Classifications can be wrong in more ways than one. As we have noted, reforms that reduce false positives—convicting the innocent—may increase false negatives—failing to convict the guilty. As usual in this area, we can only guess at the effects of this tradeoff, but the adversarial nature of criminal litigation makes it much more complicated. Everybody in an intensive-care unit—doctor, nurse, or technician—has the same objectives: the survival and health of the patient. In court, the defendant and his lawyer do what they can to undermine the work of the prosecutor and the police—to get a dismissal or an acquittal—whether the defendant is innocent or guilty.

And then there is the question of cost. The American medical system is famously well funded. It accounts for 16% of our gross domestic product. There are, of course, huge problems of inefficiency, lack of access, and uneven distribution of medical services, but they occur in an overall context of adequate, if not excessive, funding. The criminal justice system is starved. Few cases get anything like the attention they deserve. A plausible reform, like providing trials to 25% of felony defendants, is unattainable, and even basic good practice—for example, collecting and preserving all physical evidence in all felony cases—cannot be done on existing budgets.[9]

The Production of Evidence

When the wrong person is arrested, prosecuted, and convicted, it usually means that the evidence against

him was defective. The most important kinds of evidence for the prosecution are eyewitness testimony about what was done and who did it; physical evidence such as fingerprints, DNA, or stolen goods; and confessions. Most reforms designed to reduce the number of false convictions involve improving the collection, interpretation, and preservation of these kinds of evidence. That applies even when the focus seems to be elsewhere. For example, careful scrutiny of jailhouse snitches is important, in large part because they generally claim to report confessions by defendants, and pseudoscientific expertise, such as handwriting analysis, can provide dangerously misleading interpretations of critical items of physical evidence.

To maximize the amount of high-quality evidence, investigations should be scrupulous and thorough, even when the case against a suspect already seems to be convincing. This is most obvious with regard to physical evidence such as fingerprints, blood, semen, surveillance tapes, weapons, and other objects related to the crime. Many physical traces are ephemeral. Rain obliterates footprints, friends carry off incriminating objects, the scene of the crime is compromised, and evidence that could throw light on the crime is irretrievably lost. It is crucial that the initial search be comprehensive—rather than focused exclusively on collecting evidence against the identified suspect—and the evidence that is collected should be carefully preserved for future analysis. If DNA testing of critical evidence is possible, it should always be done. Forensic testing should be done in laboratories that are held to high standards, operate independently from police departments, and are regularly monitored (National Research Council, 2009). Unfortunately, many American crime labs fall far short of this ideal. All this will cost money, but it would be money well spent since it would increase the likelihood both of finding the true criminal in the first place and of discovering mistakes after the fact.

The use of DNA identification in rape cases illustrates the benefits of careful attention to physical evidence. Twenty-five years ago, a rape trial in which the defendant claimed to be misidentified was usually a battle of credibility: the jury had to decide whose story to believe, the victim's or the defendant's. Now, if semen is recovered, DNA testing decides most of these cases, and they rarely go to trial. And in old cases, an innocent man serving time for rape may be exonerated, and the real rapist may be identified, by comparing the sample to profiles in DNA databases—but only if semen from the crime scene was collected and preserved. In the years to come, new technologies may extend this scenario to other tests and other crimes, if the collection and preservation of the physical evidence is conscientious.

In principle, the same logic applies to interrogations, eyewitness testimony, and physical evidence that cannot be tested by means as definitive as DNA identification. If an interrogation is recorded and the recording is preserved, it is easier to tell whether the incriminating facts were provided by the suspect or by the interrogator. Recording interrogations may reduce false confessions because the police will be less likely to coerce or mislead the suspect if they know that the defense attorney and possibly the judge or jury will be able to see how the confession was obtained. If, later on, new evidence suggests that a defendant who was convicted on the basis of a confession might be innocent, the tape can be reviewed in order to reassess the authenticity of the confession.

In order to eliminate intentional or inadvertent suggestive police pressure on eyewitnesses, the officer who conducts the lineup should not know which person is the actual suspect. Several other procedures that can improve the accuracy of lineup identifications are currently used by some police departments. First, the other people in the lineup are chosen on the basis of the witness's description of the suspect, making sure that the suspect has no identifying feature that makes him stand out: a person who did not witness the crime but who read the witness's initial description should not be able to pick out the suspect (Doob and Kirschenbaum, 1973). Second, the witness is told that the criminal may not be in the lineup. Third, as soon as the witness has made a choice, she is asked how confident she is about that choice (cf. Wells et al., 1998). Fourth, if there are several witnesses, they are shown the lineup one at a time, with no information about how the others responded. All of these are good practices, and future technology may provide further improvements. For example, it may be possible to create a photo lineup on a laptop soon after a possible suspect is apprehended and show it to witnesses while their memories are still fresh.

Finally, as with police interrogations, video recording the identification procedures may inhibit police bias at the time of the identification and will create a record that can be reviewed in case of later doubts about its accuracy. Recordings of interrogations and identifications will rarely provide evidence as strong as a DNA sample, but they are far better than what we have now—inconsistent recollections of police, suspects, and witnesses.

Like extra care in collecting and preserving physical evidence, these reforms will cost money. There are other costs as well. A clear DNA match or mismatch does not raise the problem of false negatives, of letting guilty people go free. With these less conclusive forms of evidence, the very tactics that lead to false convictions may increase the number of true convictions,

and preventing the police from using these tactics will likely reduce the number of true convictions. Misleading a guilty suspect about the strength of the evidence against him may induce him to confess. Directing a witness's attention to the suspect in the lineup or urging her to make an identification may give her the confidence to identify the guilty person. Reporting an ambiguous fingerprint as a clear match might provide the extra evidence necessary to secure the conviction of the true criminal.

Recordings of interrogations or lineups may also provide powerful ammunition for shrewd defense attorneys, who could peruse them for any irregularities that may raise questions in the mind of the judge or jury, even if these irregularities should seem trivial in the context of the whole procedure. That is an inevitable consequence of the adversarial system and probably the major reason that police so often resist proposed reforms.

Big Cases and Small Cases

Almost all of the false convictions we know about—those that end in exoneration—are big cases: murders and rapes for which innocent defendants were convicted at trial and sentenced to death, life imprisonment, or decades behind bars. A case of this scope consumes hundreds or thousands of hours of effort by police officers and lawyers on both sides. Big cases are fertile ground for confirmation bias: there are many stages, many pressures, and many opportunities for investigators to become committed to their theories. Perhaps as a result, these cases also frequently involve serious misconduct by the attorneys or the officers involved. The most common type of government misconduct that we know about is the suppression of exculpatory evidence (Armstrong and Possley, 1999), but some cases include perjury by police officers (for example, forensic analysts), procuring perjury by civilian witnesses, and planting physical evidence (Gross et al., 2005). When such misconduct is discovered, it is rarely punished (Ridolfi, 2007). On the defense side, the main failing is incompetence—lawyers who do nothing to prepare for trial, never talk to their clients, or ignore alibi witnesses and exculpatory physical evidence. Here, too, the rules are unenforced (Possley and Seargeant, 2011). Even egregious neglect rarely results in reversals of convictions or sanctions against the offending lawyer.

Addressing the problems of big cases is comparatively straightforward, at least in the abstract. They are already time-consuming, uncommon, expensive enterprises, and it would not take much more time and money to do things right. Government misconduct and incompetent defense should not be tolerated.[10] It would not take a substantial increase in resources to collect and preserve physical evidence, conduct careful identification procedures, record interrogations, or conduct systematic internal review within prosecution and police agencies to identify investigative errors before trial. O'Brien (2009) found that confirmation bias was greatly reduced if the subjects were asked to list evidence against, as well as in favor of, their theory of the case. Perhaps some version of that procedure would reduce false convictions, or a prosecutor or a police officer with no other role in the investigation could review the case as a devil's advocate, looking for unexplored theories and evidence of possible errors (see also Findley and Scott, 2006).

The overwhelming majority of all criminal convictions, however, are comparatively small, routine cases: guilty pleas after cursory investigations. In the usual case, nobody—neither the defense nor the prosecution, and certainly not the court—collects any evidence once charges have been filed; as a practical matter, the initial police report, however sketchy, forms the only factual basis for a negotiated plea bargain. Some of these cases may involve affirmative misconduct—perjury, intimidation, concealing exculpatory evidence—but the nearly universal problem is simply inattention. An innocent defendant in a small case is likely to have two unattractive choices: take a bargain and plead guilty or hold out for trial, perhaps in pretrial custody, and hope that by then someone will come up with evidence of his innocence.

Inevitably, most false convictions happen in small cases, but we very rarely spot them. A global reform of plea bargaining in ordinary cases—for example, requiring an independent factual investigation by the defense attorney—would involve a basic restructuring of the system of criminal litigation and a huge infusion of money. Some reform of this sort might be worth the cost, but it is unlikely to happen in the foreseeable future and we do not know enough about false convictions in run-of-the-mill cases to know what sort of change is most likely to help. Eliminating plea bargaining entirely and providing trials to all or most defendants is out of reach, and there is no reason to believe that doing so would improve the accuracy of convictions. The alternative to a guilty plea is usually a trial, and the main reason that innocent defendants plead guilty is fear that they might be convicted at trial and receive much longer sentences. In most cases that fear is probably justified. For example, of the 35 defendants in the Tulia mass exoneration, 8 went to trial, were convicted of drug dealing, and received sentences that averaged nearly 47 years and ranged up to life imprisonment. The other 27 Tulia defendants pled guilty: 1 was not sentenced, 11 received some combination of probationary terms and fines, and 15

were sentenced to terms that averaged about 7 years (Gross, 2008).

Our only suggestion for preventing false convictions in comparatively small criminal cases is the most basic and amorphous: those who handle such cases should remain alert to the possibility that the defendant might be innocent. This applies to everyone, from police officers to judges, but it is especially important for defense attorneys, who have unlimited access to the defendants and whose job it is to protect them.

Conclusion

This chapter began with a famous quotation from Judge Learned Hand. As we conclude, it may be instructive to read it again, but in the context in which it was written (*United States v. Garsson*, 1923). The question before the court was whether the defendant was entitled to see the evidence considered by the grand jury that indicted him. Judge Hand held that he was not:

> Under our criminal procedure the accused has every advantage. While the prosecution is held rigidly to the charge, he need not disclose the barest outline of his defense. He is immune from question or comment on his silence; he cannot be convicted when there is the least fair doubt in the minds of any one of the twelve. Why in addition he should in advance have the whole evidence against him to pick over at his leisure, and make his defense, fairly or foully, I have never been able to see. No doubt grand juries err and indictments are calamities to honest men, but we must work with human beings and we can correct such errors only at too large a price. Our dangers do not lie in too little tenderness to the accused. Our procedure has been always haunted by the ghost of the innocent man convicted. It is an unreal dream. What we need to fear is the archaic formalism and the watery sentiment that obstructs, delays, and defeats the prosecution of crime.

In short, procedures that help criminal defendants are far more likely to obstruct the conviction of the guilty than to protect the innocent. On the specific issue that Judge Hand decided, his argument is unconvincing. In most states, grand jury records are now routinely turned over to defendants, along with many other types of prosecutorial evidence, with no apparent harm. But the fear that Hand expressed remains a basic argument against many possible reforms.

Sometimes (as with grand jury records) this reaction is nothing more than anxiety about change. Many police chiefs, for example, complain in advance that if they are required to record all station-house interrogations, there will be a steep drop off in confessions and convictions; but in jurisdictions where this rule is implemented, the police soon switch sides and become advocates for recording (Sullivan, 2004). On other issues the problem is more complicated.

In theory, we guarantee every indigent criminal defendant an effective legal defense at state expense. But if we actually provided high-quality defense in every case (and we do not, not nearly), it would be harder to get convictions. Defense lawyers who actually investigate their cases will spot some false charges, but more often they will make the state work harder to convict the guilty. The state may have to find more evidence, do more legal work, and perhaps take more cases to trial rather than resolve them with guilty pleas. Even if the defense attorneys do not succeed in getting acquittals or dismissals for their guilty clients, the prosecutors and the police will have less time to pursue other criminals. That is Judge Hand's basic complaint.

Extreme versions of this argument are ugly. It may be cheap to convict defendants by manufacturing perjured evidence, or there may be no other way to nail a murderer you *know* is guilty, but nobody advocates perjury as a policy. On more mundane issues, however—conducting thorough investigations, providing effective defense attorneys, disclosing evidence that is unfavorable to the state, there is a serious problem. Our criminal justice system cannot possibly function as the rules say it is supposed to with the funds that we provide. Instead, we take shortcuts, of which the most common is plea bargaining, which papers over all holes in the work that precedes the guilty plea. If we actually require our public servants to do careful work, many fewer crimes will be prosecuted, unless we also greatly increase their budgets. Police and prosecutors must be forgiven for not believing that any increase in the work demanded of them will be matched by an increase in funding.

There are more than a million felony convictions a year in the United States, mostly for property or drug offenses, and millions of misdemeanor convictions. The sentences most defendants receive are comparatively light, but only comparatively. A year in jail is a harsh punishment by ordinary standards, and arrest, pretrial detention, and criminal conviction are severe punishments in themselves even if there is no post-trial incarceration. The laboratory research on factors that increase or decrease false convictions is irrelevant to most of these cases. There is often no eyewitness other than the arresting officer, no lineup, no formal interrogation. In some small cases the suspect is innocent, but our knowledge is so limited that we can offer little in the way of recommendation except to say that the problem of false convictions in this context is potentially very serious and deserves research.

Our main suggestion is distressingly vague: everyone involved in processing such routine criminals should be on the lookout for cases of possible innocence.

For major crimes, especially the murders and rapes that dominate known exonerations, we have mentioned a variety of possible reforms throughout this chapter. Most are costly, but we believe that they are worth the money. We will not achieve accuracy, either in identifying and convicting criminals or in protecting innocent suspects, by continuing to give in to our penchant for handling criminal investigations and prosecutions on the cheap.

In a world of adequate funding, we would simply say that the police and the lawyers should do what they are supposed to do and follow the practices we and others recommend. In the system that exists, we need to set priorities. We see two, and they bracket the criminal process:

First, if the initial investigation by the police is careless or incomplete, information is lost forever. Physical evidence that is lost or destroyed cannot be replaced. An interrogation that is not recorded cannot be reconstructed. Eyewitness memory that is altered by a suggestive lineup or suggestive questioning cannot be retrieved. All of these steps happen before any defense investigation can possibly begin. That means that the state has a critical responsibility to collect and preserve physical evidence, record interrogations, and conduct and record careful nonsuggestive eyewitness identifications.

Second, we should be less rigid about reopening criminal cases after conviction. No legal system can function if court judgments are subject to open-ended review, but that principle has limits. It is uncommon for substantial evidence of innocence to emerge after conviction, but when that happens, there is a real possibility that the defendant is innocent. The most efficient way to limit the harm caused by convicting the innocent is to reconsider convictions with an open mind when new evidence calls them into doubt, rather than reject the possibility because it is too late.

Notes

1. The case of Gary Dodson, who was exonerated in Illinois in August 1989 (Connors et al., 1996), is sometimes mistakenly described as the first DNA exoneration in the United States (e.g., Gross et al., 2005).

2. Unless we specify that we are discussing mass exonerations, we use the term *exoneration* to refer to cases of innocent defendants who were released as a result of proceedings that affected only their individual cases.

3. Our definition of *exoneration* also excludes known defendants who are almost certainly innocent but who have not been *exonerated*—frequently because they pled guilty to reduced charges in order to obtain freedom. For example, in 1978 Terry Harrington and Curtis McGhee were convicted of murder in Iowa. In 2003, twenty-five years later, the Iowa Supreme Court reversed the convictions because the police had concealed evidence about another suspect. By then all the key prosecution witnesses had recanted their testimony. Both defendants were offered a deal: plead guilty to second-degree murder and go free. Harrington turned down the deal, and charges were later dismissed after the state's star witness recanted once more; he was exonerated. McGhee decided to play it safe, took the deal, and was released. He does not count as exonerated since the final outcome of his case was a conviction, even though he is just as likely to be innocent as his codefendant (Gross et al., 2005).

4. Some researchers have attempted to estimate the rate of false convictions indirectly. Huff et al. (1996) surveyed officials who work in the criminal justice system and report that the great majority believe that wrongful convictions are rare—in the range of 1%. As Gross and O'Brien (2008) pointed out, that estimate is just collective guess work—and self-serving optimism to boot. Poveda (2001) tried to balance Huff's low estimate with data from surveys of prisoners, about 15% of whom claim to be innocent, but two unreliable and biased estimates are no better than one. Other researchers have used statistical models that build on the frequency of disagreements on verdicts between trial judges and juries, as reflected in surveys of criminal trial judges, to estimate that up to 10% of criminal convictions in jury trials are erroneous (Gastwirth and Sinclair, 1998; Spencer, 2007). These models, however, do not.

5. As Gross and O'Brien (2008) pointed out, most death-sentenced inmates are removed from death row and resentenced to life imprisonment, frequently within a few years of conviction, after which they are unlikely to receive the extraordinary attention and scrutiny that are devoted to reinvestigating and reviewing the cases of prisoners who may be put to death. And, of course, some false convictions must remain undetected even for defendants who are executed or die on death row from other causes.

6. A defendant who pleads guilty may also have the right to appeal, but the appeal is usually limited to procedural issues that concern the entry of the guilty plea or the legality of the sentence.

7. The effect of appellate review may be much greater among capital cases, where the rate of reversal of death *sentences*, if not the underlying convictions, is far higher than the reversal rate for any other category of criminal judgments (Liebman et al., 2000). If judges are more likely to reverse death sentences when they think the defendant may be innocent—and there is strong anecdotal evidence to that effect—this would mean that most innocent capital defendants are removed from death row for procedural reasons even if they are not exonerated.

8. We are aware of a couple of recent attempts to open the process of factual review in adversarial systems of litigation but have insufficient information to evaluate their

efficacy: (1) In 1997, Great Britain, which has an adversarial common-law system that is similar in many respects to that in the United States, created a Criminal Cases Review Commission, which has the power to investigate complaints by prisoners that they were wrongfully convicted and to refer claims it deems meritorious to the appellate courts. In its first ten years, the courts took action on 313 referrals from the commission and exonerated the defendants in 187 cases, 68% of those referred (Criminal Cases Review Commission, 2009). (2) In 2007, the State of North Carolina created an Innocence Inquiry Commission that has some of the features of the British Criminal Cases Review Commission (North Carolina Innocence Inquiry Commission, 2009).

9. The federal government is an exception. The federal criminal justice system is far better financed than the state systems, from investigative agencies and prosecutors through defense attorneys and courts. There are very few exonerations in federal cases, which might in part reflect the impact of better funding, but federal cases differ sharply from state cases in many other respects as well. For example, federal cases account for about 6% of felony convictions and about 12.5% of prison inmates, but only about 1.7% of convicted murderers are in federal prisons, and murder cases account for the majority of all exonerations in the past 30 years.

10. Part of the reason for lax enforcement of the professional rules against prosecutorial misconduct and defense attorney incompetence is the belief by courts and disciplinary authorities that defendants are guilty, so no harm, no foul. The defendants usually are guilty, but that is no justification for ignoring constitutional requirements and rules of professional conduct. One way or the other, enforcing these rules cannot depend on discovering miscarriages of justice. Most are never detected, and even when they are, the time lag is so long that the offending attorney has probably forgotten all about it, or has retired, or died—or become a judge.

References

Armstrong, K., and Possley, M. (1999, January 10–14). Trial and error. How prosecutors sacrifice justice to win. *Chicago Tribune*.

Bedau, H. A., and Radelet, M. L. (1987). Miscarriages of justice in potentially capital cases. *Stanford Law Review*, *40*, 21–179.

Borchard, E. M. (1932). *Convicting the innocent: Errors of criminal justice*. New Haven: Yale University Press.

Cohen, T. H., and Reeves, B. A. (2006). *Felony defendants in large urban counties, 2002* (NCJ 210818). Washington, DC: U.S. Department of Justice, Bureau of Justice Statistics.

Connery, D. (1977). *Guilty until proven innocent*. New York: G. P. Putnam's Sons.

Connors, E., Lundgren, T., Miller, N., and McEwen, T. (1996). *Convicted by juries, exonerated by science: Case studies in the use of DNA evidence to establish innocence after trial* (NCJ 161258). Washington, DC: U.S. Department of Justice, National Institute of Justice.

Criminal Cases Review Commission. (2009). *Annual Report and Accounts 2008/09*. London: The Stationery Office. Retrieved from http://www.ccrc.gov.uk/CCRC_Uploads/ANNUAL_REPORT_AND_ACCOUNTS_2008_9.pdf

Damaska, M. R. (1986). *The faces of justice and state authority*. New Haven, CT: Yale University Press.

Davis, T. Y. (1982). Affirmed: A study of criminal appeals and decision-making norms in a California court of appeal. *American Bar Foundation Research Journal*, *7*, 543–648.

Death Penalty Information Center. (2009). Innocence and the death penalty. Retrieved from http://www.deathpenaltyinfo.org/innocence-and-death-penalty

Doob, A. N., and Kirschenbaum, H. (1973). Bias in police lineups—Partial remembering. *Journal of Police Science and Administration*, *1*, 287–293.

Drizin, S. A., and Leo, R. (2004). The problem of false confessions in the post-DNA world. *North Carolina Law Review*, *82*(3), 891–1007.

Durose, M. R., and Langan, P. A. (2003). *Felony sentences in state courts, 2000* (NCJ 198821). Washington, DC: U.S. Department of Justice, Bureau of Justice Statistics.

———. (2004). *Felony sentences in state courts, 2002* (NCJ 206916). Washington, DC: U.S. Department of Justice, Bureau of Justice Statistics.

———. (2007). *Felony sentences in state courts, 2004* (NCJ 215646). Washington, DC: U.S. Department of Justice, Bureau of Justice Statistics.

Findley, K. A., and Scott, M. S. (2006). The multiple dimensions of tunnel vision in criminal cases. *Wisconsin Law Review*, *2*, 291–397.

Frank, J., and Frank, B. (1957). *Not guilty*. Garden City, N.Y.: Doubleday.

Frisbie, T., and Garrett, R. (2005). *Victims of justice revisited*. Evanston, IL: Northwestern University Press.

Fuller, L. (1961). The adversary system. In H. Berman (Ed.), *Talks on American law* (pp. 10–22). New York: Vintage Books.

Furman v. Georgia, 408 U.S. 238 (1972).

Gardner, E. S. (1952). *The court of last resort*. New York: William Sloane Associates.

Garrett, B. (2008). Judging innocence. *Columbia Law Review*, *108*, 55–142.

Garrett, B., and Neufeld, P. (2009). Invalid forensic science testimony and wrongful convictions. *Virginia Law Review*, *95*, 1–97.

Gastwirth, J. L., and Sinclair, M. D. (1998). Diagnostic test methodology in the design and analysis of judge-jury agreement studies, *Jurimetrics*, *39*, 59–78.

Givelber, D. (2001). The adversary system and historical accuracy: Can we do better? In S. D. Westervelt and J. A. Humphrey (Eds.), *Wrongly convicted: Perspectives on failed justice* (pp. 253–268). Piscataway, NJ: Rutgers University Press.

Gregg v. Georgia, 428 U.S. 153 (1976).

Gross, S. R. (1987). Loss of innocence: Eyewitness identification and proof of guilt. *Journal of Legal Studies*, 16, 395–453.

———. (1998). Lost lives: miscarriages of justice in capital cases. *Law and Contemporary Problems*, *61*(4), 125–152.

———. (2008). Convicting the innocent. *Annual Review of Law and Social Science*, *4*, 173–92.

Gross, S. R., Jacoby, K., Matheson, D. J., Montgomery, N., and Patil, S. (2005). Exonerations in the United States 1989 through 2003. *Journal of Criminal Law and Criminology*, *95*, 523–560.

Gross, S. L., and O'Brien, B. (2008). Frequency and predictors of false conviction: Why we know so little, and new data on capital cases. *Journal of Empirical Legal Studies*, *5*, 927–962.

Heron, M. (2007). Deaths: Leading causes for 2004. *National Vital Statistics Reports*, *56*(5), 1–96.

Hoffman, M. (2007, April 26). The 'innocence' myth. *Wall Street Journal*, p. A19.

Huff, C. R., Rattner, A., and Sagarin, E. (1996). *Convicted but innocent: Wrongful conviction and public policy.* Thousand Oaks, CA: Sage.

Innocence Project. (2009). Know the cases. Retrieved from http://www.innocenceproject.org/know/Browse-Profiles.php

Kansas v. Marsh, 126 S. Ct. 2516 (June 26, 2006).

Kassin, S. (2005). On the psychology of confessions: Does innocence put innocents at risk? *American Psychologist*, *60*, 215–228.

———. (2008). False confessions: Causes, consequences, and implications for reform. *Current Directions in Psychological Science*, *17*, 249–253.

Leo, R. (2008). *Police interrogation and American justice.* Cambridge, MA: Harvard University Press.

Leo, R. A. (2009). False confessions: Causes, consequences, and implications. *Journal of the American Academy of Psychiatry and the Law*, *37*(3), 332–343.

Liebman, J. S., Fagan, J., and West, V. (2000). A broken system: Error rates in capital cases, 1973–1995. Retrieved from http://www2.law.columbia.edu/instructionalservices/liebman/

Liptak, A. (2008, March 25). Consensus on counting the innocent: We can't. *New York Times.* Retrieved from http://www.nytimes.com/2008/03/25/us/25bar.html

Lofquist, W. S. (2001). Whodunit? An examination of the production of wrongful convictions. In S. D. Westervelt and J. A. Humphrey (Eds.), *Wrongly convicted: Perspectives on failed justice* (pp. 253–268). Piscataway, NJ: Rutgers University Press.

Mathieson, A., and Gross, S. R. (2004). Review for error. *Law, Probability and Risk*, *2*, 259–268.

McGonigle, S., and Emily, J. (2008, October 10). 18 Dallas County cases overturned by DNA relied heavily on eyewitness testimony. *Dallas Morning News.*

Meissner, C. A., and Brigham, J. C. (2001). Thirty years of investigating own-race bias in memory for faces: A meta-analysis. *Psychology, Public Policy, and Law*, *7*, 3–35.

National Research Council. (2009). *Strengthening forensic science in the United States: A path forward.* Washington D.C.: The National Academy Press.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175–220.

North Carolina Innocence Inquiry Commission (2009). Retrieved from http://www.innocencecommission-nc.gov/

O'Brien, B. (2009). A Recipe for bias: An empirical look at the interplay between institutional incentives and bounded rationality in prosecutorial decision making. *Missouri Law Review*, *74*, 999–1050.

PBS. (2004, June 17). *The Plea.* Frontline. Retrieved from http://www.pbs.org/wgbh/pages/frontline/shows/plea/four/stewart.html

Pennington, N., and Hastie, R. (1992). Explaining the evidence: Tests of the story model for juror decision making. *Journal of Personality and Social Psychology*, *62*, 189–206.

Possley, M., and Seargeant, J. (2011). Preventable error: Prosecutorial misconduct in California 2010. Northern California Innocence Project. A Veritas Initiative Report. Retrieved from http://www.veritasinitiative.org/wp-content/uploads/2011/03/Prosecutorial Misconduct_FirstAnnual_Final8.pdf

Poveda, T. G. (2001). Estimating wrongful convictions. *Justice Quarterly*, *18*(3), 698–708.

Pronovost, P., Needham, D., Berenholtz, S., Sinopoli, D., Chu, H., Cosgrove, S., et al. (2006). An intervention to decrease catheter-related bloodstream infections in the ICU. *New England Journal of Medicine*, *355*, 2725–2732. Retrieved from http://content.nejm.org/cgi/content/full/355/26/2725

Radelet, M. L., Bedau, H. A., and Putnam, C. (1992). *In spite of innocence.* Boston: Northeastern University Press.

Radin, E. D. (1964). *The innocents.* New York: Morrow.

Ridolfi, K. (2007). *Prosecutorial misconduct: A systematic review.* Preliminary Report prepared for the California Commission on the Fair Administration of Justice.

Risinger, D. M. (2007). Innocents convicted: An empirically justified factual wrongful conviction rate. *Journal of Criminal Law and Criminology*, *97*, 761–806.

Ross, L. D., and Nisbett, R. (1991) *The person and the situation*. New York: McGraw Hill College.

Scalia, J. (2001). *Federal criminal appeals, 1999, with trends 1985–99* (NCJ 185055). Washington, DC: U.S. Department of Justice, Bureau of Justice Statistics.

Scheck, B., Neufeld, P., and Dwyer, J. (2003). *Actual innocence: When justice goes wrong and how to make it right*. New York: Signet.

Schlup v. Delo, 513 U.S. 298 (1995).

Spencer, B. D. (2007). Estimating the accuracy of jury verdicts, *Journal of Empirical Legal Studies*, *4*, 305–329.

Strier, F. (1996). Making jury trials more truthful. *University of California at Davis Law Review*, *30*, 142–151.

Sullivan, T. P. (2004). *Police experiences with recording custodial interrogations*. Report presented by Northwestern School of Law, Center on Wrongful Convictions. Retrieved from http://www.jenner.com/system/assets/publications/7965/original/CWC_article_with_Index.final.pdf?1324498948

Tavris, C., and Aronson, E. (2007). *Mistakes were made (but not by me)*. New York: Harcourt.

Thibaut, J., and Walker, L. (1975). *Procedural justice: A psychological analysis*. Mahwah, NJ: Lawrence Erlbaum.

United States v. Garsson 291 F. 646 (L. Hand J.) (1923).

Vaughan, D. (1996). *The Challenger launch decision*. Chicago: University of Chicago Press.

Warden, R. (2004). The snitch system. Chicago: Northwestern School of Law, Center on Wrongful Convictions. Retrieved from http://www.law.northwestern.edu/wrongfulconvictions/issues/causesandremedies/snitches/SnitchSystemBooklet.pdf

Wells, G. L. (1978). Applied eyewitness testimony variables: System variables and estimator variables. *Journal of Personality and Social Psychology*, *36*, 1546–1557.

———. (2006). *Does the sequential lineup reduce accurate identification in addition to reducing mistaken identification?* Retrieved from www.psychology.iastate.edu/faculty/gwells/SequentialNotesonlossofhits.htm

Wells, G. L., and Bradfield, A. L. (1998) "Good, you identified the suspect": Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology*, *83*, 360–376.

Wells, G. L., Lindsay, R. C., and Ferguson, T. J. (1979). Accuracy, confidence, and juror perceptions in eyewitness identification. *Journal of Applied Psychology*, *64*(4), 440–448. doi:10.1037/0021-9010.64.4.440

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., and Brionacombe, C.A.E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads, *Law and Human Behanvior*, *22*, 603–647.

Wells, T., and Leo, R. A. (2008). *The wrong guys: murder, false confessions and the Norfolk Four*. New York: The New Press.

CHAPTER 10

Behavioral Issues of Punishment, Retribution, and Deterrence

JOHN M. DARLEY

ADAM L. ALTER

How should policy makers approach the complex issues that arise when a society attempts to minimize the "negative" behaviors of its citizens? Generally, some behaviors are deemed so harmful to society that they warrant a sanctioning response from societal agents. In the absence of official sanctions, the individuals who are harmed might attempt to administer idiosyncratic forms of justice, so any society must tackle the question of how to deal with such actions. Two significant questions arise: First, what sort of control system will capture citizens' shared moral perspective on which acts should be punished and how harshly? Second, given what we know about human cognition and behavior, what sort of control system will produce the lowest crime rates?

We will deal with these two questions in turn. First, we will review recent research on people's perceptions of wrongful actions and appropriate punishments for those actions. We will suggest that those judgments are generally intuitive rather than the product of more formal reasoning systems and that they are driven by information relevant to what punishment the offender "justly deserves" for his or her offense. We will suggest that citizens' perceptions of justice place limits on the types of societal punishment practices that will be perceived as fair and suggest that legal codes that violate those constraints cause citizens to lose respect for the justice system. We will then turn to the second question, which concerns the crime-controlling efficacy of the punishment practices our society has adopted. We will focus our attention on whether our societal control systems are optimal, or even effective, for reducing crime rates in our society and conclude that they are largely ineffective. Our general claim here is that conventional approaches to dealing with crime, punishment, and deterrence in the legislative policy arenas deviate from what research on behavioral decision making has recently discovered about how people actually think and behave.

Punishment Judgments are Just-Deserts Intuitions

We will make two basic arguments in this section. First, people's punishment decisions are based on intuitions rather than systematic reasoning processes. Second, the variable that drives people's punishment decisions is an emotionally tinged reaction of moral outrage or anger that tracks the perceived moral "weight" of the transgression. Put differently, people make punishment decisions based on what they intuitively believe the offender justly deserves.

Punishment Decisions are Intuitive

A considerable body of evidence from both laboratory and field studies suggests that punishment decisions are driven by intuition rather than systematic reasoning processes. These decisions are intuitive because they rely on heuristic processes rather than on carefully deliberated reasoning (Kahneman, 2003). Frederick and Kahneman (2002) illuminated two such heuristic processes when they reanalyzed data from an earlier study (Kahneman, Schkade, and Sunstein, 1998) in which participants from the Texas jury roll suggested punitive damages for mock civil cases. They found that the participants awarded damages in amounts that were highly correlated with a product of their outrage in response to the offending conduct and the degree of perceived harm resulting therefrom (r = .90 when the offending firm was large; r = .94 when the firm was medium sized). Frederick and Kahneman "did not intend to suggest that respondents separately assessed outrageousness and harm and then computed their product; rather, [they proposed] that the feeling of outrage is ordinarily sensitive to both the recklessness of the culprit and the suffering of the victim" (2002, p. 64). In

short, people punish using an intuitive combination of wrongfulness and harmfulness.

Other studies have similarly shown that people rely chiefly on wrongfulness when calculating appropriate sentences for criminal acts. In one study (Alter, Kernochan, and Darley, 2007b, Study 1), participants suggested appropriate sentences for shooting, theft, and assault crimes that varied according to whether they were both harmful and wrongful (an intended and completed offense), only wrongful (a failed attempt), or only harmful (a mistake). Although both harmfulness and wrongfulness mattered, wrongfulness was a significantly stronger predictor of sentence severity for all three crimes. In a second study (Alter, Kernochan, and Darley, 2007b, Study 2), Princeton University students believed they were punishing their peers for violations of the university's honor code. Again, although harmfulness and wrongfulness both mattered, students suggested sanctions that were tied most strongly to the wrongfulness of the violation. These effects also persist in the realm of criminal defenses, where people are more willing to give a dispositionally moral defendant the benefit of the *ignorance of the law* defense (Alter, Kernochan, and Darley, 2007a).

The intuitive origins of justice are nowhere more evident than in children, who evince an increasingly sophisticated tendency to punish and allocate resources as they mature. Notably, even young children who cannot explain the *reasoning* behind their judgments demonstrate remarkably robust *intuitions* (e.g., Darley and Schultz, 1990; Kohlberg, 1981; Piaget, 1932/1965). Kohlberg (1981) mapped out a series of moral stages through which every child passes in a set order. Although research has cast some doubt on the universalism of Kohlberg's theory (e.g., Gilligan, 1982; Turiel, 1983), researchers broadly agree that children develop intuitive notions of right and wrong, and justice and injustice, well before they have developed comparable logical reasoning capacity.

At the other end of the expertise spectrum, judges also demonstrate the use of common-sense intuitions in rendering punishments (Hogarth, 1971). One consequence of relying on intuition is that judges, who have widely varying political and ideological views, generate untenably inconsistent sentences for the same crimes (e.g., Tetlock et al., 2007). Legal bodies in the United States responded to such variable sentences by devising various model penal codes that were designed, in part, to standardize sentencing (Kadish, 1999). Foreign judiciaries have shown a similar tendency to rely on intuition. Negligence law in England and Australia depends largely on tests of "commonsense" (Vinogradoff, 1975), and the chief justice in the Supreme Court of New South Wales, Australia, has written *guideline judgments* that provide steps in an effort to dissuade free-form, intuitive approaches to sentencing (e.g., *R. v. Jurisic*, 1998).

Several other psychological researchers have concluded that moral decisions in general tend to be products of intuition rather than complex cognitive reasoning. They report that, at least sometimes, people make quick judgments that certain acts are clearly morally wrong but are unable to articulate the reasoning that led them to that decision. Haidt's well-known work (2001) on moral dumbfounding demonstrates that people quickly respond that, for instance, an act of brother-sister sexual intercourse is wrong. But when Haidt's experimenter pressed them to supply reasons why it is wrong, they struggled to find a logical basis for their moral position. For example, many cite reasons such as the psychic damage that could be done to one of the participants or that a genetically defective baby could result, even though those reasons cannot be true, as the experimenter pointed out, because elements in the story told to the participants ruled out those possibilities. Having run out of reasons why it is wrong, the participants often ended by asserting, "it is just wrong," thus falling back on their intuitions.

Several researchers who have worked with "trolley car" problems have reported a similar effect. The core of the trolley-car scenario, invented by the philosopher Phillippa Foot, describes a runaway trolley car that, if it continues, will kill five people working on the track farther down the incline. This outcome can be changed if the respondent throws a switch to divert the trolley to another track, in which case it will kill only one person. In an alternative scenario, the respondent must fatally push a bystander from a footbridge onto the track below to save the five workers. The utilitarian considerations are identical here in the two cases: killing one to save many. Most people, however, report that although throwing the switch would be the moral action and they would do it, pushing the person onto the tracks would be immoral and they would not do it. When one set of researchers (Hauser et al., 2007), who had given their respondents both of these scenarios, buried among many others, confronted their respondents with the discrepancy, about 70% of the respondents could not provide an explanation of why they differed on the two cases, echoing Haidt's moral dumbfounding work. Based on his work and his review of other work, Haidt concluded that "moral judgments" derive from "quick automatic evaluations (intuitions)" (2001, p. 814), and Hauser et al. suggested that "much of our knowledge of morality is . . . intuitive, based on unconscious and inaccessible principles" (p. 125).

Punishment Intuitions Are Just-Deserts Based

These intuitive punishment judgments are driven by information about the moral wrongness of a harmful action. The just-deserts punishment notion is a retributive one that was articulated by Immanuel Kant, who argued that perpetrators ought to be punished in proportion to the moral offensiveness of their action. The critical just-deserts variable is the moral wrongness of the crime. That is, more morally wrong crimes deserve more serious punishments. Typically, crimes against persons are regarded as the most wrong, and they are perceived to require the most severe punishments. Further, factors that mitigate or exacerbate the morality of the action also affect punishment judgments. Thus, holding the severity of the crime constant—say, an embezzlement of a constant amount of money—one would punish less severely if the money was needed for a medical operation for the embezzler's child, and more severely if the money was needed to continue the embezzler's life of debauchery. Similarly, any circumstance that mitigates responsibility for the action, as J. L. Austin (1956) famously pointed out, would also mitigate the moral severity of the offense. In short, a person seeking retributive justice via just-deserts punishments will be concerned with those factors that exacerbate or mitigate the moral outrage provoked by the action and seek to balance moral magnitude of the offense with the punishment.

In contrast to just-deserts considerations would be the claim that punishment judgments are based on one or more utilitarian considerations, such as the desire to deter the punished offender from future offenses (specific deterrence), the desire to signal to the general populace that the action in question is forbidden (general deterrence), or the desire to lock away a dangerous person who will commit future dangerous actions (incapacitation).

One way of ascertaining people's punishment goals would be simply to ask them. Generally when this is done, respondents tend to agree with statements representing aspects of all three of those goals. As we have argued, however, the sorts of punishment judgments made in response to the presentation of a specific case are intuitive judgments. The workings of the intuitive system are automatic and not recoverable by introspection, a case made most powerfully by Nisbett and Wilson (1977), who first demonstrated that in a series of judgment tasks that we now would say were of the sort that are made intuitively, the actual decisions made depended on certain factors, such as the positioning of the chosen object in the choice array. When asked why they made a particular choice, participants appeared to formulate an after-the-fact answer that did not correspond to the experimentally demonstrated actual basis for their choice. Second, and probably equally influentially, Nisbett and Wilson reviewed several other areas of research in which the real determinants of the respondents' judgments were not recoverable by the respondents. Since their paper appeared, psychologists have been skeptical about relying on the veridicality of people's reports about their internal decision processes.

Happily, another method for discovering the determinants of people's judgments exists. In fact, it was the method that Nisbett and Wilson used to discover the real determinants of their subjects' choices. This approach, which is frequently described as a policy-capturing approach (Cooksey, 1996), is employed in punishment research as follows. First, the researcher thinks through the factors that are important to know if one is determining a punishment for an offense based on, say, a retributive stance, but which are of less importance from an incapacitative stance, then she considers next which factors make a difference from, say, a deterrence perspective, and so on. For instance, any information about whether the offender is likely to reoffend is mainly important to those who seek to sentence in order to protect society from habitual criminals, an incapacitative stance.

The first study in the sequence (Darley, Carlsmith, and Robinson, 2000) will illustrate this method for discovering the actual determinants of punishment judgments. Participants in this study read ten vignettes that described various crimes ranging from quite minor (e.g., stealing music CDs from a store) to quite severe (e.g., political assassination). The vignettes also varied the prior history of the perpetrator: he was described either as a first-time offender with no history of misbehavior or as a repeat offender who had committed similar crimes in the past. In this way, Darley, Carlsmith, and Robinson manipulated the moral severity of the crime (a retributive factor) and the likelihood of future offenses (an incapacitative factor). The dependent variable was the duration of the recommended prison term.

The results were quite clear. When it came to the duration of the sentence, people were highly sensitive to the severity of the offense and largely ignored the likelihood that the person would offend again. Manipulation checks revealed that people clearly understood whether the perpetrators were likely to reoffend but did not use that knowledge in determining sentence severity. Rather, it was the just-deserts considerations that determined the punishment severities assigned. In short, people punished exclusively in proportion to the moral weight of the crime.

One other result of the study tends to confirm the central role of just deserts in punishment decisions. After making their initial ratings (the ones we have just discussed), the participants were asked to review the vignettes a second and a third time from retributive and incapacitative perspectives. We explained to participants in some detail the goals of these theories, how they differed from each other, and how they operated in general. We then asked them to assign punishment again from these perspectives.

People adopting the retributive perspective gave sentences that tracked the moral severity of the offense closely and did not track the incapacitative factors. When they adopted the utilitarian, incapacitative perspective, they became highly sensitive to the future risk of the offenders, as one would expect, but neither did they ignore the moral severity of the offense. The moral severity of the crime intruded on their sentencing and remained a significant predictor of the sentence. This suggests the persistent importance of just-deserts factors in the punishment decisions of the ordinary person.

Furthermore, the punishment decisions made in the subjects' first pass through the cases, when they were making their own decisions uninstructed by the experiments, strongly resembled the decisions they made when they were instructed to make decisions based on a just-deserts stance and were quite different from the decisions made when instructed to use an incapacitative stance. Again, these findings suggest that the just-deserts, retributive, stance is the one people normally take when they are punishing moral offenders.

In a second series of studies, Carlsmith, Darley, and Robinson (2002) assessed whether retribution or deterrence considerations more strongly motivate people to punish. Although participants claimed that both retribution and deterrence were important factors, their actual sentences consistently reflected retributive goals to the exclusion of deterrence goals. In each of three studies, Carlsmith, Darley, and Robinson manipulated separately the harmfulness and moral gravity of each wrongful action. For example, in one study, the perpetrator either embezzled money from his firm (low deservingness) or dumped toxic chemicals in a public waterway (high deservingness). Each crime also varied according to whether it was easy to detect (low need for deterrence) or difficult to detect (high need for deterrence). Although participants' suggested sentences were highly attuned to the moral gravity of the crime (retribution), they almost entirely disregarded how easily the crime could be detected (deterrence). A mediational analysis showed that moral outrage, the emotional reaction presumed to accompany retribution, mediated the relationship between punishment deservingness and punishment severity. Like Darley et al. (2000), Carlsmith, Darley, and Robinson (2002) managed to isolate the determinant of people's sentencing decisions by manipulating each of the potential determinants orthogonally. In both cases, participants punished offenders who behaved immorally and largely ignored, or paid less attention to, deterrence, incapacitation, and other pragmatic concerns.

Punishment Decisions as Just-Deserts Intuitions: Policy Consequences for Legal Codes

By now, it should be clear that there is considerable evidence that people are intuitive retributionists. They punish offenders according to how strongly those offenders deserve punishment (e.g., Carlsmith, Darley, and Robinson, 2002) and quantify deservingness chiefly according to how wrongfully the offender behaves (Alter, Kernochan, and Darley, 2007b). The question then arises of whether and to what extent code drafters should consider these intuitive legal codes in the minds of the populace when enacting laws. One argument for doing so is that in a democratic society, the views of the citizens should at least be considered by code drafters. Code drafters may disagree with some of the views of the citizens, but they should at least seek to persuade the citizens of the moral correctness of the code drafters' stance.

There is a strong policy argument for, in so far as possible, shaping the general patterns of the criminal codes around the shared moral codes of the citizens. As psychological researchers have begun to examine the determinants of law-abidingness, they have discovered that the more citizens regard laws as "moral," the more they are likely to obey them. Tyler's major study (1990) on why people obey the law is generally recognized as the beginning of the demonstration of the importance of procedural justice in promoting law-abidingness, but it also presented compelling evidence of the importance of perceptions that the law is morally right in promoting law-abidingness. He reviewed (p. 37) four previous studies that found correlations between the perceived morality of a law and the degree to which the perceiver has obeyed the law in the past and intends to obey it in the future. In his own large-scale panel study on compliance with the law, the perceived moral appropriateness of the laws made an important contribution to predicting people's willingness to obey the laws (p. 60).

Developing Disrespect for the Law

Several studies have examined the contention that is most relevant to our argument—that the perception that the laws are "immoral" causes an unwillingness to

obey the laws. Two studies (Silberman, 1976; Tuttle, 1980) reported correlational results indicating that those who regard the laws as immoral are less likely to obey them. More recently, psychologists have sought to demonstrate that respondents who discover that the legal system strays from their moral sensibilities move toward disrespect for and disobedience of the law. Greene (2003) and Nadler (2005) independently showed that people develop contempt for such legal systems and become willing to engage in minor legal violations that, in their minds, restore the balance of fairness. Greene asked participants to read newspaper articles that reported manifestly unfair punishments; in some cases people were punished too harshly (e.g., a couple sentenced to five years in prison for engaging in oral sex contrary to an Idaho statute) and in others, not severely enough (e.g., a teenager is exonerated after refusing to intervene when his friend rapes a young girl). Relative to those who read articles describing fair punishments, participants who read unfair cases tended to report greater dissatisfaction with the law and a greater desire to violate minor laws. Similarly, in Nadler's study, participants read articles that described soon-to-be-introduced legislation that was either fair or unfair. Later, participants who read articles outlining unfair legislation scored higher on a "likelihood of criminal behavior" questionnaire.

Unfair legal systems also tend to encourage jury nullification (Greene, 2003; Nadler, 2005), in which jurors refuse to return guilty verdicts against defendants who unequivocally satisfy the legal tests for conviction. Nullification pushes the democratic nature of jury trials to its limits, and regular displays of nullification are a clear indicator that the legal system fails to reflect prevailing community mores. For example, nullification was particularly prominent during the early Civil Rights era, when all-white juries regularly refused to convict white defendants who killed black victims (Conrad, 1998). In one study (Nadler, 2005), some participants read the story of David Cash, a homeless defendant who stole a shopping cart and was sentenced to life imprisonment under a "three strikes" mandatory sentencing law. Understandably irate, these participants were more likely to acquit a plainly guilty defendant in the second phase of the study.

What we have made here is a utilitarian argument for having the criminal justice system organized around the one nonutilitarian principle for distributing punishment—just deserts. The argument is that distributing punishments in ways that clash with the moral sensibilities of citizens erodes their willingness to obey the law. At the policy level, it is worth asking whether this is a concern that needs to be addressed. Is there a gap between legal codes and community sentiments?

The Increasing Gap between Codes and Community Sentiments

The answer is yes. There are a number of gaps, and scholars suggest that such gaps are growing. Unfortunately, the legal system at large is moving progressively further from people's sensibilities (Kadish, 1999). American legislatures are steadily increasing the number of acts that attract criminal sentences, many of them trivial and prosecuted capriciously (Coffee, 1991, 1992). According to California law, for example, landowners can be punished severely for allowing wastage of artesian water on their property (California Water Code, 2003). Coffee (1992) has even suggested that the criminal law has begun to encroach on civil law territory, so that civil wrongs that once attracted civil damages are now punished with criminal fines and even prison sentences. More and more offenses are strict liability offenses, for which no showing of a deliberate commission of a known moral wrong is required. This draconian approach to criminal prosecution and sentencing offends the morality-based just deserts approach that we suggest people endorse. As such, the criminal legal system is increasingly likely to attract contempt, flouting, and jury nullification and, at a second-order level, risk losing its moral credibility as a legitimate source of guidance as to whether actions are right or wrong.

On the Chances of Achieving Deterrence by Increasing Sentence Duration

We will next examine an actual practice of the criminal justice system from the perspective of recent discoveries in behavioral research and suggest that the practice is close to useless. The policy in question is the system's increasing reliance on lengthy prison sentences to deter crime.

Since the 1960s, politicians in the United States have legislated considerable increases in the duration of the prison sentences assigned for a good many crimes. American crime-control policy makers have sought to reduce or even reverse, what was perceived to be frightening increases in crime rates by greatly increasing the standard penalties assigned to most crimes. Policy makers have thus sought to deter crimes by manipulating one of the three determinants of the deterrent weight of the punishment: specifically, they have chosen to increase *sentence severity* as opposed to increasing the probability of detection and conviction for the crime or shortening the interval between the commission of the criminal act and the eventual imprisonment for the crime.

Is this likely to be an effective way of reducing crime? We first review the aggregated studies that gather evidence on the degree to which upward adjustments in the duration of a criminal penalty cause reductions in the aggregated rate of commission of that crime. This has been taken as the appropriate measure of whether lengthening the duration of sentences accomplishes an increase in the deterrent force of the penalty. Examining the conclusions of the available criminological reviews of the issue, we conclude that it has not. We will next examine why it has not been effective. Drawing on recently developed behavioral-science thinking about what we might call the psychology of the decision making of a person contemplating a crime, we will suggest why it is that manipulations of sentence duration will be ineffective in controlling crime rates.

It is important to be precise about exactly what we are claiming about the relationship between the canonical existence of criminal punishments, generally prison terms, and the rates of crime commission. We do not doubt that the generalized knowledge that penalties follow if one commits actions that the society considers criminal does influence the conduct of society members. We note that a consistently present institution in many, perhaps all, societies, is a "criminal justice system" to inflict these punishments. Here we are echoing the conclusion of Nagin (1998), one of the leading empirical researchers on the control of criminal behavior, "that criminal punishment has had deterrent effects." What we do suggest is that several decades of legislative activity that has increased piecemeal the duration of the sentence for any crime that is currently attracting the public's attention will not be effective in reducing the aggregate rate of that crime.

Second, we will suggest that the recently accumulated evidence about how individuals make decisions reveals why lengthening the duration of sentences is generally an ineffective deterrent. We do not think that gains in crime control are impossible from a behavioral-decision-making point of view, but we think that increasing sentence duration is not the way to get them. We suggest ways in which a deterrent force can successfully be brought to bear on crime commission, largely through increasing the salience of the surveillance mechanism.

Finally, the policy implications of these arguments will be considered. We are paying the grim societal price associated with incarcerating more and more people for longer and longer prison sentences without reaping the gains that were assumed to follow from these practices.

The United States Project: Increasing Sentence Duration to Deter

Our society seems to have converged on the following causal story about crime and its control. We believe in agentic actors, that is, actors who are the origins of their own actions and who exert a high degree of control over the commission of these criminal actions. Furthermore, we hold the theory of control that is thereby implied, which is that if enough punishment follows the action in question, then the actor will be deterred from the commission of the action. Moreover, imprisonment fulfills the goal of incapacitating the criminal while he or she undergoes the reformation process associated with appropriately swift and severe punishment. The task of our criminal justice system is to administer that punishment, to deter the actor from the commission of those crimes. This practice has the useful side effect of signaling to other actors who are contemplating crimes that they should refrain from committing them. This is the general deterrence effect. All of this was usefully formalized by Jeremy Bentham (1830/1975), who suggested that the task of the criminal justice system was to set the penalty for a crime high enough so that it would outweigh whatever gains or pleasures that the criminal would achieve by committing the crime. In other words, our culture tends to think about crimes from a rational-actor perspective, although Becker (1968) and others have found it useful to reinstate the power of this perspective from time to time.

Since the 1960s, the United States criminal justice system has been engaged in a policy of attempting to deter crime commission by increasing the duration of the sentence assigned to the crime. Criminal codes are generally set by state legislatures, and states have rivaled each other in being "tough on crime." It may well be that our legislatures are filled with ardent Benthamites, attempting to manipulate the deterrence calculus, although a public-choice perspective would suggest that legislators are instead responding to rational calculations about being voted out of office if they leave open any possibility of being accused of "being soft on crime" by rival candidates.

Whatever the reasons for this expansion, it has swelled the prison populations. In 2004, 726 of every 100,000 residents were incarcerated, the highest rate in U.S. history and the highest rate in the Western world (Barkow, 2006). Meanwhile, sentences have continued to lengthen, partially due to legislatively imposed increases in sentences for most offenses beginning about 1972. There has also been a pattern of "upgrading" offenses in criminal codes, such as upgrading drug possession offenses from misdemeanors

to felonies. Finally, many states have passed "three strikes and you're out" laws which assign startlingly long sentences to felons who have committed a third offense. And furthermore, many states have enacted truth-in-sentencing laws that, for instance, deny any possibility of parole until quite high a percentage of the prison sentence has actually been served. The net result has been to increase the prison population of the United States, which was estimated at around 600,000 in 1972, to well over 2 million persons in the first years of the twenty-first century (Tonry, 2004). Given that prison budgets come from the states' discretionary, rather than mandated, funds, there has been some crowding out of expenditures on other systems, such as school systems, by the increased costs needed for the state penitentiary systems to cope with this rise in the incarcerated populations.

Looking back, it is worth noting that when legislators decided to increase sentence severity, they faced two general options for doing so: either making the moment-by-moment experience of suffering in prison more dire, or attempting to increase the negativity of the prison experience by increasing its duration. By and large, legislative decision makers have used the latter technique. Some legislatures, however, have made decisions that were conceived as increasing the negativity of the prison experience. This included reducing or eliminating prison educational programs, exercise facilities, libraries, televisions, and other entertainment facilities so as to diminish the "country club" perceptions of prisons. And many public-opinion observers have perhaps sensed, rather than measured, the willingness of segments of the population to mandate physical inflictions of pain for criminal offenders.

Clearly, it would be possible to increase the severity of punishments more directly. In prior times, societies quite actively varied the severity of punishments, ranging from torture to floggings to cruel inflictions of the death penalty. For a complex set of reasons, best explicated by David Garland (1990), at least Western societies no longer engage in these inflictions. And those reading this almost certainly agree with this renunciation. One criminal justice scholar does not. Graeme Newman published a book in 1985 titled *Just and painful: A case for the corporal punishment of criminals*, in which he advocated a sort of electronic flogging. His comment was that this was likely to be a more effective deterrent, and his case was that it did not inflict the usual horrible consequences of long prison terms on criminals. In the prologue to the second edition (1995), he reported the overwhelming negative responses his book got from the media, a response that suggests that we will not adopt such practices. So, by and large, we have varied our inflictions of what Bentham called the severity of the "pain" on criminals by adjustments in the duration of prison sentences (until those increases are insufficient and we resort to the death penalty). By this approach, if we seek to affect the deterrent effect of the penalty by increasing the severity of it, we will accomplish this by manipulations of sentence duration.

Are Increases in Sentence Duration Effective in Reducing Rates of Crime?

Have the quite-considerable increases in sentence severity achieved corresponding decreases in the rates of commission of various crimes? There are two recent reviews of this question. The first, by von Hirsch and his associates (1999), is rather cautious in its conclusions. To begin with, the authors are rather tactful: "The evidence concerning severity effects is less impressive" (p. 47). As they go on, they reveal that they do not find it impressive at all: "Present association research, mirroring earlier studies, fails—as just noted—to disclose significant and consistent negative associations between severity levels and crime rates" (p. 47).

Doob and Webster (2003) came to a more forthright conclusion: "Based on the weight of the evidence, including recent evidence made available from 'three strikes' laws, we will not obtain general deterrence effects by alterations in sentence severity that are within the limits that are plausible in Western countries" (p. 143).

Researchers like von Hirsch reveal a note of surprise that evidence is not found for what must seem like a tautology—the more severe the sentence, the more deterrent effect it must have. This may account for the general impulse of researchers to continue to look for severity-produced crime-rate reduction and the continued use by politicians of the strategy of increasing sentence severity to control what are perceived as rising crime rates.

The Effects of Probabilistic Future Outcomes on Decisions in the Present

The behavioral-decision-making community, we suspect, is not at all surprised by the empirical evidence, because they are aware of a number of reasons why a legislatively imposed upward change in the sentence duration for some crimes does not "make its way" through the multiple steps necessary to affect the behavior of potential offenders. The primary reason is that regardless of how we manipulate severity,

we need to remember, as Bentham pointed out, that there are two other components of punishment which can be expected to influence its efficacy as a deterrent: first, the *certainty* with which it is administered, and second, the *celerity* of its administration. It is useful to equate certainty with the likelihood that a crime committed will be detected and the criminal found, tried, convicted, sentenced, and imprisoned. The celerity with which the punishment is administered might refer to the interval from the commission of the crime to the convicted criminal entering the prison to serve his sentence; alternatively, given human representational capabilities, it could begin when the offender is arrested and charged with the offense. What this means is that the deterrent weight of a punishment is some function of the severity of the punishment, the likelihood of receiving the punishment, and the delay with which it is either pronounced or administered after the offense. Of course, celerity and certainty are more likely to influence a potential offender's conduct when he or she has offended previously. Specifically, the difference between experiencing punishment as swift and consistent, or slow and erratic, is likely to be more immediate and behavior-altering for a person who has experienced these consequences directly rather than learning of them from others.

Robinson and Darley (2004) considered the likelihood that an offender would be identified and ultimately imprisoned, and the average length of time between the commission of the crime and his or her imprisonment. They found that of all offenses that are reported as committed on the various surveys of citizens, only 1.3% result in the identification of the criminal and subsequent punishment and that only 1 of every 100 of those convicted is sent to prison. The gap between the commission of a crime and imprisonment is also quite long (if the person is convicted at all). Even defendants who plead guilty experience a mean arrest-to-sentencing delay of 7.2 months, and those who elect not to plead guilty wait an average of 12.6 months. The question, then, is how these gaps influence the deterrent impact of criminal sentences. It is usual to think of these functions as multiplicative, but this need not be true of its general shape. It is, of course, likely to be true when there is no possibility of being detected or when the severity of the punishment is so mild as to be negligible. That is, if one is highly unlikely to be apprehended, then, intuitively, even severe punishments could be risked; or if the punishment is mild, even if it is likely to occur, it could be risked. Given (as reported above) that conviction probabilities for most criminal acts are 1.3% or below and the delays between crime and (low probability) punishments often exceed a year, punishments are unlikely to be so severe as to raise the deterrent weight of the function to affect criminal behavior significantly.

The interval from arrest to conviction to prison, the "celerity" term in Bentham's equation, has not been the subject of much analysis, perhaps because court dockets are notoriously crowded and delays are often sought by one or the other side in the American adversarial system. But that dearth of study is regrettable, because evidence from one of the rare celerity studies suggests that it may be a powerful variable when the particularities of the criminal justice system make possible experimenting with its full range. Interestingly, many legislatures have moved toward allowing summary punishments, in which the crime is punished on detection and detection quickly follows the crime. Using interrupted time-series models, Wagenaar and Maldonado-Molina (2007) analyzed the effects of changes in state laws that now allow for immediate administrative (i.e., preconviction) suspension of driving licenses of drivers who failed a breath test when they were pulled over. They report that these policies "have significant and substantively important effects in reducing alcohol-related fatal crash involvement by 5%, representing at least 800 lives saved per year in the United States." The effect was seen at all drinking levels, from below the legal limits to extremely drunk drivers. "In clear contrast, post conviction license suspension policies have no discernable effects." They also draw the lesson that "penalties delayed, even if relatively severe, do not have clearly demonstrable effects on behavior. But penalties applied immediately, even if more modest, have clear deterrent effects."

The Behavioral Perspective: Time Discounting

Perhaps the best-known discovery of researchers of behavioral decision making is that events that will or may happen in the future exert less sway on decisions in the present than they rationally should. This has been referred to as *hyperbolic discounting*, but more recently, with the realization that the discounting function is not always hyperbolic in shape, the phenomenon has been referred to by the more neutral term of *time discounting* (e.g., Read, 2001).

As should be clear, time discounting is highly relevant to punishment policies. Although popular legal television shows convey the impression that sentencing snaps at the heels of arrest, the actual delay between arrest and sentencing is substantial. In the United States, the arrest-sentencing lag exceeds the duration of many sentences, ranging from 3 months in Seattle, Washington, to 15 months in Hackensack, New Jersey, with the median lag exceeding 6 months across all U.S. criminal courts (Ostrom and Kauder, 1999).

The delay between arrest and sentencing distinguishes criminal activity from other forms of labor. Whereas professionals and tradesmen cannot enjoy financial rewards before they earn qualifications and a reputation, criminals enjoy their ill-gotten gains today, and only occasionally suffer punishment later (Davis, 1988; Wilson and Herrnstein, 1985). This cost-following-benefit reversal, where the cost of punishment is not a certainty, has important consequences for the justice system. Just as many Americans forego a four-year bachelor's degree because the distant promise of financial reward is both remote and uncertain, would-be criminals impulsively seize a criminal opportunity because the prospect of retribution is too distant and insufficiently certain to deter. Again, although would-be criminals are likely to learn of the typical delay between crime and punishment from other offenders, the effects of time discounting might apply more directly to repeat offenders.

The powerful tendency to discount the punitive force of distant punishments (or rewards), known as time discounting (e.g., Chung and Herrnstein, 1967; Loewenstein and Prelec, 1992), is best illustrated by contrasting two hypothetical legal systems. Suppose a person were deciding whether to rob a house, an offense that carries a one-year prison sentence. Under the present legal system, the sentence will not be imposed for six months, so the would-be criminal is forced to imagine the severity of a year in prison beginning in six months. Psychologically, the offender retains his or her freedom over the coming six months. Now, suppose a utopian legal system were significantly more efficient such that the one-year sentence began the day after arrest. The offender is now forced to imagine the prospect of a prison sentence beginning tomorrow, rather than in six months.

For a number of reasons, a swiftly imposed sentence feels intuitively more severe. First, it is easier to imagine the negative consequences of losing one's freedom tomorrow than in six months. An image of *tomorrow* is already partially formed, both building on activities that exist today and functioning as the starting point for new activities. In contrast, a 24-hour period in six months is laced with uncertainty, so it seems less valuable. Supporting this claim, researchers have shown that students are willing to donate 85 minutes to help struggling peers during a distant midterm period, but only 27 minutes during the coming week (Pronin, Olivola, and Kennedy, 2008). Importantly, students donated only 46 minutes of time during the midterm period when the experimenter reminded them that they would probably face similar time pressures as they did presently. This final condition suggests that people can be induced to think more deeply about the value of their time in the future, although their natural tendency is to value their time more highly in the present than in the future.

A second factor that weakens the force of future punishments is the difficulty of imagining oneself in the future. One consequence of this difficulty is that people often imagine their future selves as different people altogether (Nussbaum, Trope, and Liberman, 2003; Pronin and Ross, 2006). Naturally, the perceived force of a sentence is lessened when a diluted version of the self seems to bear the burden.

A third factor, the tendency to represent future events more abstractly than present events (known as Construal Level Theory, Trope and Liberman, 2003), also suggests that distant punishment weighs less heavily than imminent punishment. According to Construal Level Theory, people construe present events at a relatively specific, concrete level, and future events at a relatively general, abstract level. A sentence beginning today might therefore loom larger, replete with specific images of an imposing cellmate, abject boredom, and confinement in a tiny prison cell. In contrast, imprisonment conjures a much vaguer image when viewed from afar, represented by the abstract and therefore less vivid concepts that accompany incarceration and the loss of freedom.

The Behavioral Perspective: Duration Neglect

When fines and community service sentences are too lenient, judges face the difficult task of selecting an appropriately lengthy prison sentence. The naive logic governing imprisonment duration is simple: the punishment should reflect the gravity of the crime, so longer prison sentences are more severe than shorter sentences. Thus, a 10-year sentence should have 10 times the punitive "bite" of a 1-year sentence (Robinson and Darley, 2004). Consequently, when considering whether or not to reoffend, a potential offender should remember the longer sentence as more aversive and should therefore be less likely to reoffend than the offender who served a shorter sentence. This approach theoretically satisfies the sentencing goals of retribution, incapacitation, and deterrence, because the trauma of a longer sentence should deter greater offenders more strongly than a shorter sentence deters lesser offenders.

A Salience-Based System of Deterrence

During a debate before the Guatemalan presidential elections of 2007, Patriotic Party candidate Otto Perez Molina stared into the camera and said, "I am addressing the criminals that I want to talk directly

to. I know some of you are watching me." In truth, Molina was speaking to the Guatemalan voters when he vowed to "increase the size of the police force by 50 percent and revive the death penalty" (Zacharia, 2007). In response, his chief opponent, Alvaro Colom of the National Unity of Hope, promised to overhaul the security services and judicial system. This sort of tough-on-crime one-upmanship is popular among politicians and rests on the flawed assumption that harsher sentences will deter crime. If harsher sentences generally have little effect on crime rates, deterrence seems like an unlikely preventive goal, certainly less so than the reactive goals of retribution and incapacitation. However, harsher sentences address only one limb of a two-pronged approach. Rather than focusing on *severity*, we believe that deterrence advocates would do better to concentrate on the *certainty* and *celerity*, or sureness and swiftness, of punishment.

Reviewing the Arguments

An increasing pool of evidence suggests that the punishment decisions of people, at least in the United States, are made intuitively and are based on just-deserts considerations. We suggest that to the extent that these culturally shared judgments of the citizens are contradicted by the legal codes and are inevitably brought to the attention of citizens, citizens may lose respect for the moral credibility of legal codes and cease to take them as trustworthy sources of guidance on how to behave. Furthermore, we suggest that, in fact, the legal codes have increasingly deviated from citizen intuitions and thus are courting the dangers we point out.

References

Alter, A. L., Kernochan, J., and Darley, J. M (2007a). Morality influences how people apply the ignorance of the law defense. *Law and Society Review*, *41*, 819–864.

———. (2007b). Transgression wrongfulness outweighs its harmfulness as a determinant of sentence severity. *Law and Human Behavior*, *31*, 319–335.

Alter, A. L., Oppenheimer, D. M., Epley, N., and Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General*, *136*, 569–576.

Anderson, D. A. (2002). The deterrence hypothesis and picking pockets at the pickpocket's hanging. *American Law and Economics Review*, *4*, 295–313.

Ariely, D., Kahneman, D., and Loewenstein, G. (2000). Joint comment on "When does duration matter in judgment and decision making?" (Ariely and Loewenstein, 2000). *Journal of Experimental Psychology: General*, *129*, 529–534.

Austin, J. L. (1956). A plea for excuses: The presidential address. *Proceedings of the Aristotelian Society* (New Series), *57*, 1–30.

Barkow, R. (2006). The political market for criminal justice. *Michigan Law Review*, *104*, 1713–1724.

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy*, *76*, 169–217.

Bentham, J. (1830/1975). *The rationale of punishment*. Montclair, NJ: Patterson Smith.

Bregy, F. A. (1901). Should the grand jury be abolished? *American Law Register*, *49*, 191–194.

Brickman, P., and Campbell, D. (1971). Hedonic relativism and planning the good society. In M. H. Appley (Ed.), *Adaptation-level theory: A symposium* (pp. 287–305). New York: Academic Press.

Bureau of Justice Statistics (2010). *Crime victimization in the United States, 2007 statistical tables* (NCJ 227669). U.S. Department of Justice. Retrieved from http://bjs.ojp.usdoj.gov/content/pub/pdf/cvus0702.pdf

California Water Code. (2003). *California laws for water wells, monitoring wells, cathodic protection wells, geothermal heat exchange wells*. Department of Water Resources, State of California. Retrieved from http://www.water.ca.gov/pubs/groundwater/california_laws_for_water_wells__monitoring_wells__cathodic_protection_wells__geothermal_heat_exchange_wells_2003/ca_water_laws_2003.pdf

Carlsmith, K. M., Darley, J. M., and Robinson, P. H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, *83*, 284–299.

Chung, S. H., and Herrnstein, R. J. (1967). Choice and delay of reinforcement. *Journal of the Experimental Analysis of Behavior*, *10*, 67–64.

Coffee, Jr., J. C. (1991). Does "unlawful" mean "criminal"?: Reflections on a disappearing tort/crime distinction in American law. *Boston University Law Review*, *71*, 193–246.

———. (1992). Paradigms lost: The blurring of the criminal and civil law models—and what can be done about it. *Yale Law Journal*, *101*, 1875–1893.

Conrad, C. S. (1998). *Jury nullification: The evolution of a doctrine*. Durham, NC: Carolina Academic Press.

Cooksey, R. W. (1996). *Judgment analysis: Theory, method, and applications*. San Diego, CA: Academic Press.

Corman, H., and Mocan, H. N. (2000). A time-series analysis of crime, deterrence, and drug abuse in New York City. *American Economic Review*, *90*, 584–604.

Darley, J. M., Carlsmith, K. M., and Robinson, P. H. (2000). Incapacitation and just deserts as motives for punishment. *Law and Human Behavior*, *24*, 659–684.

Darley, J. M., and Schultz, T. R. (1990). Moral rules: Their content and acquisition. *Annual Review of Psychology*, *41*, 525–556.

Davis, M. L. (1988). Time and punishment: An intertemporal model of crime. *Journal of Political Economy*, *96*, 383–390.

Dick, P. K. (2002). *Minority report*. London: Gollancz.

Doob, A. N., and Webster, C. M. (2003). Sentence severity and crime: accepting the null hypothesis. In M. Tonry (Ed.), *Crime and justice: A review of research* (Vol. 30, pp. 143–195). Chicago: Univ. of Chicago Press.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*, 24–42.

Frederick, S. and Kahneman, D. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, and D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). New York: Cambridge University Press.

Frederick, S. Loewenstein, G., and O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature*, *15*, 351–401.

Fredrickson, B. L. (2000). Extracting meaning from past affective experiences: The importance of peaks, ends, and specific emotions. *Cognition and Emotion*, *14*, 577–606.

Fredrickson, B. L., and Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology*, *65*, 44–55.

Garland, D. (1990). *Punishment and modern society: A study in social theory*. Oxford. Oxford University Press.

Gatrell, V.A.C. (1994). *The hanging tree*. New York: Oxford University Press.

Gilbert, D. T., Pinel, E. C., Wilson, T. D., Blumberg, S. J., and Wheatley, T. (1998). Immune neglect: A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology*, *75*, 617–638.

Gill, M., and Spriggs, A. (2005). *Assessing the impact of CCTV*. Home Office Research Study 292. Retrieved from https://www.cctvusergroup.com/downloads/file/Martin%20gill.pdf

Gilligan, C. (1982). *In a different voice: Psychological theory and women's development*. Cambridge, MA: Harvard University Press.

Gottfredson, M., and Hirschi, T. (1990). *A general theory of crime*. Stanford, Calif. : Stanford University Press.

Greene, E. (2003). Effects of disagreements between legal codes and lay intuitions on respect for the law (Unpublished doctoral dissertation). Princeton University.

Gromet, D. M., and Darley, J. M. (2006). Restoration and retribution: How including retributive components affects the acceptability of restorative justice procedures. *Social Justice Research*, *19*, 395–432.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*, 814–834.

Hauser, M., Cushman, F., Young, L., Kang-Xing Jin, R., and Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind and Language*, *22*(1), 1–21. doi:10.1111/j.1468-0017.2006.00297.x

Hochstetler, A. L. (1999). In with a bad crowd: An analysis of criminal decision-making in small groups (Unpublished doctoral dissertation). University of Tennessee, Knoxville.

Hogarth, J. (1971). *Sentencing as a human process*. Toronto, Canada: University of Toronto Press.

Kadish, S. H. (1999). Fifty years of criminal law: An opinionated review. *California Law Review*, *87*, 943–982.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist, 58*, 697–720.

Kahneman, D., Schkade, D. A., and Sunstein, C. R. (1998). Shared outrage and erratic awards: The psychology of punitive damages. *Journal of Risk and Uncertainty*, *16*, 49–86.

Kohlberg, L. (1981). *Essays on moral development (Vol. 1). The philosophy of moral development: Moral stages and the idea of justice*. San Francisco: Harper and Row.

Levitt, S. D. (2004). Understanding why crime fell in the 1990s: Four factors that explain the decline and six that do not. *Journal of Economic Perspectives*, *18*, 163–190.

Loewenstein, G., and Prelec, D. (1992). *Choices over time*. New York: Russell Sage Foundation.

Marvell, T., and Moody, C. (1996). Specification problems, police levels, and crime rates. *Criminology*, *34*, 609–646.

McGarrell, E. F., and Chermak, S. (1998). *Summary of results of JPD's 1997 directed patrol initiative*. Indianapolis, IN: Hudson Institute.

Mischel, W., Shoda, Y., and. Rodriguez, M. (1989). Delay of gratification in children. *Science*, *244*, 933–938.

Nadler, J. (2005). Flouting the law. *Texas Law Review*, *83*, 1399–1441.

Nagin, D. S. (1998). Criminal deterrence research at the onset of the twenty-first century. *Crime and Justice: A Review of Research*, *23*, 51–91.

Nagin, D. S., and Pogarsky, G. (2003). An experiment of deterrence: Cheating, self-serving bias, and impulsivity. *Criminology*, *41*, 167–191.

———. (2004). Time and punishment: Delayed consequences and criminal behavior. *Journal of Quantitative Criminology*, *20*, 295–317.

National Center on Addiction and Substance Abuse at Columbia University (1998). *CASA study shows alcohol and drugs implicated in the crimes and incarceration of 80% of men and women in prison*. Retrieved from http://www.casacolumbia.org/articlefiles/379-Behind%20Bars.pdf

Newman, G. (1995). *Just and painful: A case for the corporal punishment of criminals* (2nd ed.). Monsey, NY: Criminal Justice Press.

Nisbett, R., and Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*, 231–259.

Nussbaum, S., Trope, Y., and Liberman, N. (2003). Creeping dispositionism: The temporal dynamics of behavior prediction. *Journal of Personality and Social Psychology*, *84*, 485–497.

Ostrom, B., and Kauder, N. (1999). *Examining the work of state courts, 1998: A national perspective from the court statistics project*. Williamsburg, VA: National Center for State Courts.

Piaget, J. (1965). *The moral judgment of the child*. New York: Free Press

Pronin, E., Olivola, C. Y., and Kennedy, K. A. (2008). Doing unto future selves as you would do unto others: Psychological distance and decision making. *Personality and Social Psychology Bulletin*, *34*, 224–236.

Pronin, E., and Ross, L. (2006). Temporal differences in trait self-ascription: When the self is seen as an other. *Journal of Personality and Social Psychology*, *90*, 197–209.

R v. Jurisic (1998) 45 *New South Wales Law Reports* 209.

Read, D. (2001). Is time-discounting hyperbolic or subadditive? *Journal of Risk and Uncertainty*, *23*, 5–32.

Redelmeier, D. A., and Kahneman, D. (1996). Patient's memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain*, *116*, 3–8.

Robinson, P. H., and Darley, J. M. (2003). The role of deterrence in the formulation of criminal law rules: At its worst when doing its best. *Georgetown Law Journal*, *91*, 949–1002.

———. (2004). Does criminal law deter? A behavioral science investigation. *Oxford Journal of Legal Studies*, *24*, 173–205.

Ross, H. L., LaFree, G., and McCleary, R. (1990). Can mandatory jail laws deter drunk driving? The Arizona case. *Journal of Criminal Law and Criminology*, *81*, 156–170.

Seguin, J. R., Arseneault, L., Boulerice, B., Harden, P. W., and Tremblay, R. E. (2002). Response perseveration in adolescent boys with stable and unstable histories of physical aggression: the role of underlying processes. *Journal of Child Psychology and Psychiatry*, *43*, 481–494.

Sherman, L. W., and Rogan, D. P. (1995). Effects of gun seizures on gun violence: Hot spot patrols in Kansas City. *Justice Quarterly*, *12*, 673–693.

Silberman, M. (1976). Toward a theory of criminal deterrence. *American Sociological Review*, *41*(3), 442–461.

Tetlock, P. E., Visser, P., Singh, R., Polifroni, M., Elson, B., Mazzocco, P., and Rescober, P. (2007). People as intuitive prosecutors: The impact of social control motives on attributions of responsibility. *Journal of Experimental Social Psychology*, *43*, 195–209.

Tonry, M. (2004). *Thinking about crime: Sense and sensibility in American penal culture*. New York: Oxford University Press.

Trope, Y., and Liberman, N. (2003). Temporal construal. *Psychological Review*, *110*, 403–421.

Turiel, E. (1983). *The development of social knowledge: Morality and convention*. New York: Cambridge University Press.

Tuttle, W. M., Jr. (1980). *Race riot: Chicago in the Red Summer of 1919*. New York: Atheneum.

Tversky, A., and Kahneman, D. (1983). Extensional vs. intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.

Tyler, T. R. (1990). *Why people obey the law*. New Haven, CT: Yale University Press.

Vinogradoff, P. (1975). *Common-sense in law*. New York: Arno Press.

Von Hirsch, A., Bottoms, A., Burney, E., and Wikström, P. O. (1999). *Criminal deterrence and sentence severity*. Oxford: Hart Publishing.

Wagenaar, A. C., and Maldonado-Molina, M. M. (2007). Effects of drivers' license suspension policies on alcohol-related crash involvement: Long-term follow-up in 46 states. *Alcoholism: Clinical and Experimental Research*, *31*, 1399–1406.

Wilson, J. Q., and Herrnstein, R. J. (1985). *Crime and human nature*. New York: Simon and Schuster.

Zacharia, J. (2007, August 31). Guatemala candidates pledge crackdown on crime after elections. *Bloomberg .com Online News*. Retrieved from http://www.bloomberg.com/apps/news?pid=20601086&sid=aG7IhGVYCxcs&refer=latin_america

CHAPTER 11

Claims and Denials of Bias and Their Implications for Policy

EMILY PRONIN

KATHLEEN SCHMIDT

Objectivity is hard to find. Everyday experience is rife with examples of those around us who seem to lack it completely. We see people self-servingly take credit for collective efforts, we see them defend opinions that are biased by prejudice, and we see them allow personal self-interest to influence their desires for the "greater good." People, it seems, are susceptible to a host of biases that contaminate their perception and judgments. What is perhaps most surprising, though, is not that people are so biased but that they are so inclined to claim that they are objective. Recent years have brought forth countless examples of this phenomenon. Corporate executives have denied the role of self-serving motives in dishonest accounting practices, doctors have denied the role of financial self-interest in suboptimal patient-care decisions, employers have denied the role of sexism in gender-imbalanced hiring and promotion practices, and politicians have denied the role of ideological bias in their commitments to controversial social policies.

The biases that people display in cases like this can have costly consequences. As illustrated by the above examples, these biases can cause financial peril, compromise the quality of health care, and perpetuate social inequity. In many cases, those negative outcomes could be avoided if people were able to recognize their own commissions of bias. However, people generally have a "blind spot" for their own biases; that is, they show a broad and pervasive tendency to impute bias to others while denying it in themselves. Understanding this phenomenon can help us to devise suggestions for how policy makers and policy consumers can work toward overcoming its costly consequences.

Relevance to Policy

Commissions of bias can have serious consequences in the policy arena (e.g., Bazerman, 2002; Thompson and Nadler, 2000). This chapter concerns something different from those direct consequences of bias. It concerns the consequences of people's *perceptions* of their own and others' commissions of bias; that is, of their perception that their own judgments are relatively free of bias, whereas others' judgments are relatively susceptible to it. Consider a couple of examples taken from recent events.

A few years ago in Cincinnati, an African American man died during a violent struggle with the police at a White Castle hamburger shop. Many people, particularly members of the police force and their families, friends, and colleagues saw the incident as involving a threatening and massive man who violently resisted arrest and died in the process because he had a heart condition and was high on speed. Many other people, particularly members of the African American community, saw the incident as involving an innocent and unarmed man who was accosted without cause by a gang of racist police officers who subsequently overpowered and fatally attacked him. Following the incident, individuals from both "sides" claimed that their own perspective was the objectively accurate one and that those who defended the opposite view were biased (by racism, in-group favoritism, media-induced misconceptions, etc.). As a consequence, racial tensions erupted because each side felt angered and frustrated by the other's unwillingness to take a fair and reasonable view.

Much of the profit to be made in the pharmaceutical industry rests on individual physicians' decisions about what drugs they prescribe to their patients. For this reason, pharmaceutical companies employ legions of representatives to supply physicians with information about their latest drugs. Often that information is accompanied by a personal gift to the doctor being targeted. In some cases the gift might be something small, such as a pen or writing tablet, and in other cases it might be something grand, such as an invitation to attend an all-expenses-paid cruise where one would be educated about the relevant drug. These

gifts generally have an impact, and most doctors recognize that. Importantly, though, most doctors deny that such gifts affect their *own* patient-care decisions. As a result, they fail to shield themselves from that bias, while also feeling disenchanted with their colleagues whom they view as influenced by it. Because they assume that people are aware of bias when it is affecting them, they also tend to endorse policies that rely on individuals to recognize (and then disclose) their own cases of self-interest.

Perceptions of Bias in the Self versus Others

People have a blind spot when it comes to perceiving bias. As reviewed below (and also in table 11.1), this discrepancy occurs for a range of different biases, from those that inaccurately inflate the ego, to those that foster out-group prejudice, to those that compromise rational decision making. In each case, this bias blindness has significant relevance for policy.

Self-Enhancement Biases

Most people view themselves in a rosy light (e.g., Taylor and Brown, 1988). When they lack talent or ability, they tend not to realize their deficiencies (Dunning et al., 2003). They also tend to be overly optimistic about their future outcomes (Weinstein, 1980), and they tend to disregard evidence that threatens their self-esteem (Kunda, 1987).

While most people do not maintain the illusion that they are as brilliant as Einstein or as beautiful as a fashion model, most people do generally think that they are at least smarter and better looking than "average" (e.g., Alicke and Govorun, 2005). This enhancing form of social comparison has been demonstrated on dimensions ranging from agreeableness to zest. Importantly, people show a blind spot to this ego-inflating tendency. When people rate themselves more positively than average, they insist upon the objectivity of their ratings (Ehrlinger, Gilovich, and Ross, 2005; Pronin, Lin, and Ross, 2002). They even

Table 11.1 Real-world examples of the bias blind spot and research evidence for it across various domains of bias

Bias	Definition of bias	Real-world example of bias blind spot	Research evidence for blind spot
Self-enhancement bias	Seeing oneself in an overly positive light	People fail to recognize when they are over-estimating their abilities. As a result, they set themselves up for failure and disappoint others.	Ehrlinger, Gilovich, and Ross, 2005; Friedrich, 1996; Krueger, 1998; Pronin, Lin, and Ross, 2002
Self-serving bias (re. responsibility)	Taking credit for success and denying responsibility for failure	People fail to detect their own bias in denying responsibility for failures such as poor job performance. As a result, they feel wronged when blamed.	Kruger and Gilovich, 1999; Pronin, Lin, and Ross, 2002
Self-serving bias (re. attribution)	Viewing performance criteria as valid *only* if one excels on them	People fail to recognize their bias in evaluating performance criteria (e.g., standardized tests). As a result, they resent those who denounce criteria on which they have excelled (and they see those *others* as biased).	Pronin, Lin, and Ross, 2002; Pronin and Kugler, 2007
Self-interest bias	Judging what is fair or what is best for others according to one's own personal interests	While noting the impact of self-interest on their colleagues, people including doctors, accountants, and journalists fail to recognize the effect of gifts (and other social and monetary incentives) on their own decisions.	Dana and Loewenstein, 2003; Epley and Dunning, 2000; Heath, 1999; Miller and Ratner, 1998

Table 11.1 (*continued*)

Bias	Definition of bias	Real-world example of bias blind spot	Research evidence for blind spot
Prejudice or intergroup bias	Treating in-group members better than members of stigmatized groups or out-groups	People can show biases involving racism and sexism that they deny and are, accordingly, unwilling to try to overcome. Others may exaggerate those biases, leading to feelings of hurt and anger on both sides.	Dovidio and Gaertner, 1991, 2004; Uhlmann and Cohen, 2005; Vivian and Berkowitz, 1992
Ideological bias	Forming political views based on ideology and partisanship rather than reasoned analysis	Partisan opponents assume that their own political views are the product of objective thinking but that their opponents' views are biased by ideology. As a result, they are pessimistic about reaching a fair resolution.	Cohen, 2003; Pronin, Berger, and Molouki, 2007; Robinson et al., 1995
Hindsight bias	Failing to recognize the benefit of hindsight	People evaluating military or political disasters fail to realize that those disasters were hard to predict in advance. Because people are blind to their reliance on hindsight, they blame those in charge for their "obvious" errors.	Fischhoff, 1975
Correspondence bias	Viewing others' behavior as a reflection of their internal traits rather than their situation	People judge victims of bad circumstances as responsible for their plight. Unaware of this bias, they view victims' explanations as mere "excuses." Victims fear asking for help out of concern that others will show this bias.	Miller, Baer, and Schonberg, 1979; Pronin, Lin, and Ross, 2002; Van Boven, Kamada, and Gilovich, 1999; Van Boven et al., 2003
Anchoring bias	Making numeric judgments that are affected by salient but irrelevant or nonuseful numbers	When negotiators try to advocate for their side, they may be biased by the numbers put forth by their opponent or even by irrelevant numbers in the environment. Blind to that bias, they cannot aim to correct for it.	Wilson et al., 1996

do so after being taught about the bias and invited to acknowledge its impact on themselves (Pronin, Lin, and Ross, 2002). When people consider others' self-ratings, by contrast, they expect others to be overly positive (Friedrich, 1996). This can cause problems that are of concern for policy. People are likely to find themselves in conflict with others when they think that their own hard work, judgment, motivation, and intelligence are beyond reproach but that those with whom they are dealing somehow fall short. That scenario occurs, for example, when opposing negotiators judge themselves as more willing to be fair than their opponent, when team members judge their work ethic on a joint project as greater than their peers', and when political adversaries judge themselves as more inclined to take the moral high ground.

A classic self-enhancing bias involves people's tendency to view tests on which they perform well as good and valid and tests on which they perform poorly as bad and invalid. The policy implications of this bias are significant, given the role of formal tests in selecting among individuals for outcomes ranging from spots in elite colleges to jobs in the local fire department. If individuals denounce tests on which they perform poorly—and if they are unaware of this tendency—they are likely to view even objectively

reasonable tests as unfairly discriminatory. In an experiment by Pronin, Lin, and Ross (2002) on this topic, subjects participated two at a time, and each took a purported test of social intelligence. The pairs displayed the classic bias—those told they performed well rated the test as more valid than those told they performed poorly. When the experimenter warned the subjects that a self-enhancing bias may have influenced their judgments, they were more likely to suspect that possibility in their fellow participant than in themselves. Although the test in this experiment was fake, similar responses are likely to occur on the part of people (and groups) who do well versus poorly on high-stakes tests in the real world, such as college entrance exams. In such cases, individuals and groups who perform poorly are likely to resent the "obvious bias" on the part of high performers who seek to perpetuate the tests' gate-keeping role. Meanwhile, high performers are likely to dismiss the complaints of poor performers as reflecting obvious bias on the part of those who have a personal interest in seeing the tests' demise. Both sides are likely to be blind to the bias in their own views and thus particularly resentful of the other side's accusations (accusations which, they are likely to believe, apply much better to the accusers themselves). Such asymmetric bias perceptions could contribute to long-standing policy debates about the validity and costs versus benefits of standardized tests.

Self-Interest Biases

When laypeople think of bias, often the first thing that comes to mind is the biasing effect of self-interest. We see people's views on things ranging from smoking laws to presidential elections as guided by what serves their own self-interest. People view others as heavily biased by self-interest even when they deny that bias in themselves (Miller, 1999).

In one compelling set of experiments, Miller and Ratner (1998) asked people whether their own and others' decisions to donate blood would be influenced by economic incentives, and whether their own and others' views on insurance coverage for elective abortions would be influenced by their gender. The respondents also indicated what their actual decisions would be (i.e., whether they would donate blood or support the insurance coverage). The result was that they claimed that financial self-interest would have more of an effect on *others'* decisions than on their own. They also assumed that self-interest would have more of an effect on others' decisions than those others' self-reports indicated. For example, those who claimed that financial incentives would not affect their own decisions about whether to donate blood nevertheless predicted that those incentives would affect others' decisions (see also Goethals, 1986).

Other studies have shown similar effects. For example, Heath (1999) asked CitiBank employees how much their own and others' motivation to work hard in their careers was influenced by external incentives involving financial self-interest versus other factors (e.g., intrinsic interest in the work, pleasure of learning). The result was that the bank employees saw their coworkers as more motivated than themselves by financial self-interest. The policy implications of such asymmetries in perceptions of self-interest bias are noteworthy. Those designing incentive systems are likely to place too much weight on pleasing individuals' financial self-interest at the expense of other factors that those individuals may value more. In the case of securing blood donations, it might be better to appeal more to individuals' desire to see themselves as kind and generous. In the case of motivating employees to work hard, it might be better to appeal more to their interest in learning new skills (or in garnering the respect of their colleagues; see Tyler, this volume).

Prejudice and Group-Based Biases

People's perceptions of those around them often are influenced by those others' social categories, such as their race, gender, or political affiliation. Regardless of whether people intend it, they often display stereotypes against members of minority and stigmatized groups (Dovidio and Gaertner, 1991) and against members of groups other than their own. Researchers have found that even the flimsiest of group distinctions, such as whether people prefer one modern artist versus another, incur favoritism for in-group members at the expense of out-group members (Tajfel and Turner, 1979). Consistent with the theme of this chapter, people recognize these group-based biases in others more than in themselves.

People generally prefer members of their own race even when that prejudice lingers below their conscious awareness (Greenwald and Banaji, 1995). For example, white people's self-reports of their own racism poorly predict their actual tendency to display racial stereotypes and to unconsciously favor their own race (Dovidio and Gaertner, 1991, 2004). From a policy perspective, an important instance of group favoritism occurs in the context of employment hiring decisions and other competitive selection procedures. In that context, people have been known to construct ad hoc hiring criteria that disfavor individual applicants from stigmatized groups (Norton, Vandello, and Darley, 2004; Uhlmann and Cohen, 2005). In one experiment, Uhlmann and Cohen asked subjects to indicate whether being "streetwise" or being "formally educated" was a more important criterion for the job of police chief. Male subjects chose whichever criterion they believed was associated with the

male rather than the female candidate. Importantly, they denied the operation of this bias on their hiring preferences—even while they acknowledged that it would influence other people. Interestingly, the more objective they claimed to be, the more biased they actually were.

Another type of group-related bias involves the effect of people's political party memberships on their political attitudes. Most people believe that their views on issues ranging from foreign policy to healthcare reflect their personal analysis, values, and beliefs. But the reality is that people's political views often mimic those of their political party. For example, Cohen (2003; also Pronin, Berger, and Molouki, 2007) showed that when Democrat and Republican students read about alleged welfare reform proposals, they supported whichever proposal was allegedly backed by their own party—even when the Democrat proposal was in fact more conservative than the Republican one (e.g., when it called for fewer entitlements). The participants thought their peers would be swayed by which party backed each proposal, but they themselves denied being influenced by that factor (and instead claimed to be influenced by the proposals' content). In related research, people have been shown to view others as far more influenced than themselves by ideological bias stemming from partisan affiliations (e.g., Robinson et al., 1995; Sherman, Nelson, and Ross, 2003). It is not difficult to imagine how such asymmetries could produce meaningful social consequences. Individuals who believe that their own views reflect reasoned analysis and deeply held values are likely to have little respect or concern for the views of those whom they see as biased by "shallow" or "dogmatic" considerations such as political ideology.

Cognitive Biases

Human judgment and decision making often are subject to biases that arise not from motivational needs or prejudices, but rather from cognitive errors. Accurate judgment often is undermined by people's lack of awareness of these biases and therefore by their failure to correct for them. One such bias involves a "planning fallacy" in people's estimations of how much time it will take to complete work projects—people typically underestimate how much time they will need (Buehler, Griffin, and Ross, 1994). People are unaware of this bias in their time estimations, or they would correct for it (given the harmful costs of running out of time and having to submit poor work or to suffer grueling all-night work sessions). This bias, and people's blindness to it, causes problems not only on a small scale, such as missed work deadlines, but also on a much larger scale, where, for example, the cost of such misestimations can mean wars that cost more time, money, and lives than were ever imagined when the decision to wage them was made.

One of the most well-studied cognitive biases involves people's failure to recognize the power of the situation in influencing human behavior. When observing people's actions, we generally attribute those actions to internal traits of the actor (e.g., "He went to that movie because he likes violence and gore") rather than to aspects of the situation (e.g., "He went to that movie because the other ones were sold out"). This bias has been termed the *fundamental attribution error* (Ross, 1977) or *correspondence bias* (Gilbert and Malone, 1995; Jones and Davis, 1965). In a classic demonstration of it, Jones and Harris (1967) asked subjects to read an essay that they were told was written by a student asked to offer "a short cogent defense of Castro's Cuba." Even though subjects were explicitly told that the essay writer had been assigned this pro-Castro view, they nevertheless assumed that the writer held that position in reality. More recent research has shown that although people commit this error unknowingly, they are not ignorant of others' commissions of it. Indeed, people expect—and even overestimate—others' susceptibility to this bias (Miller, Baer, and Schonberg, 1979; Pronin, Lin, and Ross, 2002; Van Boven, Kamada, and Gilovich, 1999; Van Boven et al., 2003). The implications of this are important. Individuals are likely to be wary of introducing opinions counter to their own or to those of their in-group—even if they think those opinions are worth considering—out of concern that their own position will be incorrectly labeled. Considering the valuable role in policy debates of acknowledging the validity of the other side's views and of playing devil's advocate, this phenomenon is likely to hinder fruitful policy discussion.

Causes of the Asymmetry

The foregoing review describes people's tendency to claim personal objectivity at the same time that they recognize and even exaggerate bias in others. The discussion now turns to causes of this effect. Developing an understanding of those causes is a prerequisite for designing effective strategies for combating the bias blind spot's unfortunate consequences.

Unconscious Bias and an Introspection Illusion

Biases generally operate outside of conscious awareness (e.g., Dawson, Gilovich, and Regan 2002; Wilson, Centerbar, and Brekke, 2002). That is, people often show them without intending to or even being aware that they are doing so. When it comes to

assessing their own bias, people often fail to appreciate this simple fact. They instead overrely on their conscious knowledge about whether they have intended to be (or felt themselves being) biased. In assessing others' bias, by contrast, people generally prefer to look to those others' actions or to rely on their own theories about when people are biased. Rather than trusting others' reports of whether they *intended* to be biased or *felt* that they were biased, people look to what those others actually did (e.g., "Did he hire a long string of men for the vice-president job but never a woman?") and to their own assumptions about people's bias (e.g., "Most people think only men make good leaders."). This tendency to overrely on one's introspections while giving little credence to those of others has been termed an *introspection illusion* (e.g., Pronin, 2009).

In one experiment illustrating the impact of the introspection illusion on bias perceptions, Pronin and Kugler (2007) had students take a purported social intelligence test, told them they performed poorly, and then asked them to evaluate the quality of the test. Later, when those students were asked whether they had been biased in their evaluation of the test (consistent with a bias, their evaluations were uniformly negative), they assumed they had been unbiased since their conscious thoughts and motives yielded no signs of bias. A separate group of subjects did not take the test but instead observed a peer take it. Those observers took a different approach to assessing bias. They looked to the test takers' behavior and, in particular, whether the test takers disparaged the test right after performing poorly on it. Thus, individuals attended to their own internal motives, but to a peer's actions, in assessing bias. As discussed in a later section of this chapter ("Ethical Lapses"), this tendency to judge bias by consulting one's own introspections but others' actions can create significant problems in the policy arena. Those accused of ethical wrongdoings, such as doctors accused of sacrificing their patients' best interest in exchange for gifts from drug companies, or financial experts accused of compromising their clients' best interests in exchange for their own personal gain, may commit these ethical lapses without conscious intent. Thus, while onlookers may readily detect self-interest bias in their behavior, the actors themselves may deny it based on the apparent purity of their conscious motives.

Disagreement and Naive Realism

Individuals' failure to recognize their own biases derives in part from the nature of human perception. People generally have the feeling that their perceptions of objects and events in the world around them are accurate and direct reflections of what is true in "objective reality" (Pronin, Gilovich, and Ross, 2004; Ross and Ward, 1995). If the grass looks green to us, we believe it *is* green. Research on naive realism has described the tendency for people to make this same assumption about their higher-level judgments and opinions. Thus, if the new welfare-reform bill seems fair to us, we believe it *is* fair. Because we are shielded from the influences that nonconsciously bias us toward perceiving things in particular ways, we maintain unwarranted confidence in the directness of our perceptions.

Of course, others do not always share our perceptions. In such cases, we assume that those others must be either ill-informed or (having ruled out that possibility) incapable or unwilling to view things objectively. Experiments have demonstrated people's tendency to view those whose opinions differ from their own as influenced by biases including self-interest (Reeder et al., 2005), personal affections (Frantz, 2006), political partisanship (Cohen, 2003), and unwavering ideology (Robinson et al., 1995). For example, Reeder et al. (2005) showed that the more people disagreed with President Bush's decision to invade Iraq the more they saw that decision as biased by the president's personal self-interest.

In a series of experiments, Kennedy and Pronin (2008) examined the role of disagreement in perceptions of bias regarding the debate about affirmative action. In one study, subjects were presented with the putatively moderate position of an elite university president on that issue. The more they disagreed with that president's alleged position, the more bias they imputed to her. These results are noteworthy because the subjects all rated the same target with the same position. Thus, even though her position was fixed, participants viewed her as more biased when *their* position deviated from it. In a second experiment, the university president's putative position on affirmative action was experimentally manipulated in order to be either similar to the subjects' own position or considerably divergent from that position. Subjects saw the university president as more biased when they were led to infer that they had a large disagreement with her. Notably, the president's apparent extremity on the issue did not influence their perceptions of her bias, indicating that their perceptions of bias arose from disagreement rather than from the details of their adversary's position.

People are more convinced that their own objectivity surpasses that of others when those others disagree with them. As will be discussed later in this chapter, this phenomenon can transform simple disagreements into stubborn conflicts (Kennedy and Pronin, 2008), and it can act as a barrier to resolving conflicts that are already in place (Ross and Ward, 1995).

Self-Enhancement and the Motive to Deny Bias

A final source of the bias blind spot involves people's desire to see themselves in a positive light (Roese and Olson, 2007; Sedikides, 2007). Because of the undesirable nature of being biased, people may be motivated to deny their susceptibility to bias as a way of protecting or enhancing their self-image. Indeed, research suggests that people are more likely to deny susceptibility to biases that are relatively negative rather than positive (Pronin, Lin, and Ross, 2002).

People are particularly likely to see their personal traits and abilities in an overly positive light to the extent that circumstances are sufficiently ambiguous to allow for such enhancement (e.g., Dunning, Meyerowitz, and Holzberg, 1989). Thus, while it is difficult to self-enhance when it comes to punctuality (one either is on time or one is not), people self-enhance on traits such as generosity, friendliness, and driving ability, since those can be defined in different ways. The circumstances surrounding bias perception offer another such case of ambiguity. Because biases are difficult to prove (the person who has never hired a female vice-president might simply never have had a good one apply) and because they can be defined in multiple ways (e.g., in terms of motives versus outcomes), people often have the judgmental leeway to deny being biased without it being obvious that their denials are themselves biased.

Policy Applications: Three Case Studies

People's unwillingness or inability to recognize their own biases, even while they acknowledge to the point of overestimation others' biases, holds implications for a variety of sociopolitical concerns. Three such concerns—ethical lapses, discrimination, and conflict—are discussed below with respect to how they are affected by the bias blind spot. Understanding the effects of bias perception in these various contexts can inform the implementation of wiser and more effective policies for addressing these concerns. After reviewing these cases, we proceed to a discussion of potential solutions.

Ethical Lapses

People often encounter circumstances in which their motivation to be ethical and their motivation to serve their own self-interest are at odds. Although ethics can prevail in even the most difficult of circumstances, for example, when so-called whistle-blowers risk losing their jobs in order to expose the unethical practices of their employer, there also are many cases in which individuals succumb to self-interest. Scandals engulfing corporations such as Enron and WorldCom have illustrated the large-scale financial damage incurred by fraudulent accounting practices used to achieve personal financial gain. Those scandals have resulted in the losses of hundreds of thousands of jobs and in some of the largest bankruptcies in history. They also led to the downfall of one of the world's largest accounting firms, Arthur Andersen. That firm was forced to close when its allegedly independent auditing services were found to be biased in favor of the companies who paid them to do those audits. While onlookers saw this as an obvious case of corruption, those responsible denied being criminals. That denial, the present review suggests, reflects more than a simple desire to stay out of legal trouble. Bazerman, Moore, and their colleagues (Bazerman, Loewenstein, and Moore, 2002; Moore et al., 2006) have suggested that it reflects the fact that in some cases auditors were likely biased without being consciously aware of it. Thus, while the presence of bias may seem obvious to observers focusing on the correlation between the auditors' large paychecks and their approving audits, it may not have been obvious to the auditors themselves who were focusing on their conscious motives and intentions. How might this happen? Ethical lapses in the field of medicine provide an interesting illustration, as discussed next.

Physicians receive numerous incentives from pharmaceutical companies for recommending and prescribing treatment regimens owned by those companies. Those incentives include gifts bestowed as a means of product promotion, free meals with pharmaceutical representatives, paid travel and lodging expenses to exotic locales (for attendance at company-sponsored events), financial payments for referring patients to clinical trials, and opportunities for physicians to serve as medical consultants poised to profit from the scientific results they report. A meta-analysis by Wazana (2000) revealed that physicians typically meet with pharmaceutical representatives four times per month and receive six gifts per year. Not surprisingly, these incentives generally bias their patient-care decisions in a manner consistent with financial self-interest (Dana and Loewenstein, 2003; Wazana, 2000). Moreover, in keeping with the theme of this chapter, most physicians deny that these incentives influence their own medical practices, even though they readily recognize that influence on *other* physicians (e.g., Dana and Loewenstein, 2003; McKinney et al., 1990; Wazana, 2000).

Patients rely on their physicians to provide objective recommendations. In light of the influence of pharmaceutical companies, policy makers have been called upon to intervene in order to ensure the

integrity of that patient-doctor trust. Unfortunately, as noted by Dana and Loewenstein (2003), most regulatory interventions have been based on the flawed assumption that physician self-interest bias is the result of a conscious choice to succumb to inappropriate influence. Thus, for example, current regulations limit the size of gifts in order to curb conscious temptation, and those regulations also require physicians to disclose conflicts of interest. However, the research reviewed here suggests that limiting gift size will not decrease bias (since even small gifts can have big effects) and that mandating disclosures of conflicts of interest will not ensure such disclosures (since those conflicts often are not consciously recognized). Educational initiatives may succeed in making physicians more aware of the problem (e.g., Agrawal, Saluja, and Kaczorowski, 2004), but that awareness is likely to translate into their seeing their *colleagues* as biased rather than themselves.

Persistence of Racism and Sexism

Despite significant accomplishments in the fight against prejudice and discrimination, those ills are still observable today. Inequalities such as racial and gender gaps in wages are troubling and persistent. Racism and sexism (as well as other forms of discrimination) can result from unconscious and unintended biases; thus, their persistence can be partially attributed to people's blindness to their susceptibility to those biases. Indeed, much of modern sexism and racism is shown by people who lack conscious prejudice (Dovidio and Gaertner, 2004; Son Hing et al., 2005). Many experiments have made clear the role of automatic and nonconscious processes in producing responses that favor in-groups and disfavor both outgroups and groups subject to social stigma (e.g., Fazio and Olsen, 2003; Greenwald and Banaji, 1995). The fact that people show such biases without knowing it perpetuates prejudiced practices and also limits people's efforts to combat their own prejudiced behavior. While people are likely to see the need for reducing prejudice in society as a whole, they are likely to resist policies that would restrict their own freedom of decision making (e.g., by regulating their hiring procedures) because of their perception that they personally are not susceptible to group-based prejudice. Because observers are not likely to share individuals' confidence in their personal objectivity, social tensions are likely to arise as those accused of prejudice are likely to view those accusations not only as baseless but also as signaling the self-serving (or group-serving) bias of those voicing them.

Such a scenario occurred when the Massachusetts Institute of Technology (MIT) scrutinized its hiring and promotion practices with respect to women. Their investigation revealed considerable signs of gender disparity. Perhaps most surprisingly, the faculty at the School of Science at MIT included only 22 women out of 274 professors. Further analyses suggested signs of prejudicial treatment. According to a report issued by the school, male professors earned more money, had larger offices and lab spaces, and received more distinctions than their female counterparts (MIT, Committee on Women Faculty in the School of Science, 1999; Miller and Wilson, 1999). In the wake of this report, many wondered how such a prestigious institution could have engaged in such starkly discriminatory practices.

Nancy Hopkins, a biology professor at MIT, initiated the investigation. As a junior professor, Hopkins had perceived what she viewed as the school's general unfair treatment of women, but she thought she was an exception (Diaz-Sprague, 2003). Then, after struggling to obtain adequate lab space and facing the cancellation of her course in favor of a male colleague's course, Hopkins began to suspect that the gender bias was more pervasive than she had thought. Together with faculty women across the various science departments, she wrote a letter claiming the presence of "discrimination against women faculty" that was "largely invisible and . . . almost certainly unconscious" (Hopkins, as cited in Diaz-Sprague, 2003). The university president appointed Hopkins as the head of an investigation into possible inequities. Following the investigation, MIT increased salaries and space for women faculty and in 2004 hired its first female president. Critics of the report responded by calling its findings "bogus" and denying the presence of any bias at MIT (Leo, 2002). Consistent with the tendency for disagreement to induce people to see those on the "other side" as biased, those critics labeled Hopkins herself as biased and claimed that her involvement in the investigation served to introduce bias into it.

The MIT case helps illustrate how striking patterns of discrimination can emerge over time when specific and even slight occurrences of bias go unrecognized. Such slight occurrences often occur unintentionally and unconsciously, thus making it difficult for the relevant individuals to recognize and avoid them. This suggests that the route to overcoming the problem is likely to require the institution of formal policies, and ones that are not reliant on individual awareness of bias. Of course, a problem with instituting such policies is that the individuals in need of them are likely to be resistant because of their confidence in their own objectivity.

Even when formal policies are implemented, that implementation must be done with care in order to

avoid institutionalizing the very bias the policies are designed to overcome. For example, one sort of formal policy involves constructing fixed criteria for hiring decisions in order to avoid the potential for subjective biases to enter into the process. However, as discussed earlier, people tend to set their criteria for what would make a suitable job candidate depending on the qualifications and gender of the applicants (Uhlmann and Cohen, 2005). As a result, policies that implore people to use the same criteria across applicants can backfire—if people are allowed to select those criteria *after* first viewing the qualifications of the available applicants. This suggests the importance of establishing fixed criteria in *advance* of knowing how members of different groups stack up; otherwise, the appearance of formal criteria only will lend credibility to a biased procedure.

Modern prejudice may seem relatively harmless when compared to the more overt racism and sexism of the past. However, the consequences of more subtle forms of prejudice include restrictions on economic opportunity and other serious disadvantages that undermine equality (Dovidio and Gaertner, 2004). Confronting these problems is especially difficult because implicit prejudice cannot be unearthed via introspection. Thus, fruitful policies need to take into account people's frequent blindness to their own bias rather than assuming that such bias occurs out of a conscious motive to discriminate. Possible solutions to this problem are discussed below (in the section "Fixing the Problem").

Conflict

The tendency to see bias in others but not in oneself can play a critical role in the development and escalation of conflict. It also can prevent conflicts from being resolved once they have reached a point of tense escalation. Because people are generally confident in their own objectivity, they tend to view those who disagree with them as biased (Pronin, Gilovich, and Ross, 2004; Ross and Ward, 1995). This reasoning can unleash a spiral of conflict out of mere disagreement that proceeds roughly like this (Kennedy and Pronin, 2008, 2012): Disagreement leads people to perceive those who disagree with them as biased. That perception of bias leads people to infer that their adversaries will not be willing to act fairly and reasonably. Such an inference causes people to lose faith in the possibility of peaceful resolution of their disagreement and to instead opt for a more aggressive approach. By acting aggressively, people induce the other side to view them as biased (since those on that side assume that they are in the right and that therefore no objective person would aggress against them). Once this spiral of conflict is unleashed, resolution becomes difficult, because each side resents the other's unwillingness to put their biases aside in order to reach a fair agreement.

A useful case study of bias perception in conflict is presented by the cycle of violence involving terrorist attacks and government retaliation for those attacks (or, depending on one's perspective, unjust government action and terrorist retaliation for that action). Terrorist attacks bring to light the differences in worldview held by the groups that perpetrate those attacks versus those that are victims of them. This stark reminder of those differences can instill a desire to dominate, weaken, and even destroy those on the other side. But, it is not the experience of differences in worldview alone that leads to this desire. People may respond aggressively not only because they disagree with their adversaries but also because they view their adversaries' position as the product of biased and irrational thinking. In the many terrorism-related conflicts around the world today, a common theme is that each side tends to claim a monopoly on reason and objectivity—on seeing the past and present as they really are. Policy experts point out that even suicide terrorists are not necessarily biased and irrational even though that is how their victims often perceive them (e.g., Pape, 2005). For example, Ehud Sprinzak (2000), a former adviser to Israeli Prime Ministers Yitzhak Rabin and Shimon Peres, once claimed:

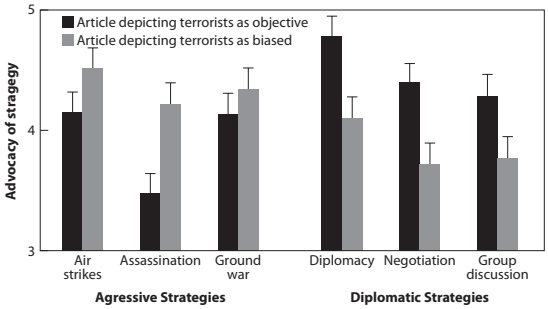
> The perception that terrorists are undeterrable fanatics who are willing to kill millions indiscriminately just to sow fear and chaos belies the reality that they are cold, rational killers who employ violence to achieve specific political objectives.

This quotation makes clear the two divergent views of terrorists. Sometimes, they are perceived as irrational fanatics biased by unbridled hatred, radical ideology, and extreme pressure to conform. Other times, they are perceived as rational warriors whose views are rooted in an objective analysis of their circumstances and of the options they think are available to them. Although both of these perspectives have been put forth in scholarly research and analysis (e.g., Crenshaw, 1998; Margalit, 2003; Merari, 2004; Pape, 2005; Post, 2005), and although the truth is probably somewhere in between, lay citizens' disagreement with terrorists' actions and beliefs typically leads them to adopt the "biased fanatic" view. Numerous studies (reviewed earlier) have shown that the more people disagree, the more they perceive those they disagree with as biased. This effect has been illustrated in the context of terrorism. One study was conducted with political activists in Northern Ireland in the wake of the Good Friday Agreement establishing the conditions

for peace in that region (described in Pronin, Lin, and Ross, 2002). Participants in the study reported that the leadership which opposed their position was especially susceptible to a host of biases that compromise fairness and objectivity in negotiation. A different study, conducted with American college students, specifically concerned suicide terrorists (Kennedy and Pronin, 2007). The more the respondents disagreed with suicide bombers, the less they viewed the bombers' actions as rooted in an objective analysis of their circumstances rather than a biased or fanatical perspective.

These perceptions of bias can perpetrate a spiral of conflict. In one experiment (Pronin, Kennedy, and Butsch, 2006), subjects were led to adopt one of two views of suicide terrorists by virtue of exposing them to an alleged *New York Times* article on the terrorist mind. Half of subjects were exposed to an article suggesting that terrorists come to their decisions via an objective analysis of the facts available to them, and the other half read an article suggesting that terrorists come to their decisions via a biased worldview. The result was that perceiving terrorists as biased versus objective powerfully affected subjects' opinions about how to combat terrorism (fig. 11.1). Those led to view terrorists as biased advocated bombing and ground attacks over negotiation and diplomacy. Those led to view terrorists as objective voiced the opposite preference.

In post–September 11 America, terrorism and the war against it are constantly subject to political debate and media sensationalism. Research on perceptions of bias makes clear that the tendency for people to view terrorists as irrational fanatics (and for terrorists to view their victims as self-serving infidels) is likely to beget a cycle of violence that cannot be easily abated. This research suggests that efforts to find diplomatic and cooperative solutions need not require both sides to see eye to eye, but rather that such efforts will require both sides to recognize that the eyes of those on the other side are no more clouded by bias than their own. If adversaries can recognize each other's potential for clear-headed thinking (as well as recognizing their own potential for biased thinking), they will be more inclined to pursue diplomatic approaches rather than viewing violence as the only option.



11.1. Advocacy of aggressive versus diplomatic strategies for combating terrorism after an article was read describing the "terrorist mind" as either rational and objective or as irrational and biased.
(Adapted from Pronin, Kennedy, and Butsch, 2006)

Fixing the Problem: From Hazardous Approaches to Promising Solutions

As illustrated by the above case studies, the policy consequences of the bias blind spot can be serious. Because biases often operate unintentionally and without awareness, averting them can be difficult. Unfortunately, many of the intuitively appealing remedies for curbing the negative effects of bias are unlikely to help and, in some cases, are likely to hurt. The remainder of this chapter will explore three prevalent and intuitively appealing approaches, with a focus on the problems with those approaches and on strategies for overcoming those problems (see table 11.2). The aim of this portion of the chapter is to offer lessons that can inform the design and implementation of policies in which bias perceptions play a role. Fortunately, a more psychologically informed perspective is likely to be effective in restricting the negative effects of the bias blind spot.

Mandating Disclosure

To the extent that individuals are inclined to deny their own biases, one obvious solution is to formally require them to openly acknowledge those biases. That solution is frequently implemented in the form of mandatory disclosure guidelines. Such guidelines are based on the premise that even if individuals do not personally have a problem with their own biases, those who interact with them may feel differently and therefore should be fully informed of such biases. Based on this premise, disclosure is one of the more common measures implemented to combat the problem of bias. Physicians are required to disclose payments received for patient referrals, stockbrokers are required to disclose if they have a financial interest in companies whose stock they recommend, and researchers are required to disclose funding sources that have a stake in the outcome of the results.

THE PROBLEM WITH MANDATING DISCLOSURE

Unfortunately, disclosure may not work for a number of reasons that primarily arise from people's lack

Table 11.2 Potential pitfalls and solutions regarding common strategies for attacking the bias blind spot

Type of solution	Basic idea	Possible pitfalls	Making it work
Mandating disclosure	Advisors disclose potential conflicts of interest.	Advisors cannot disclose conflicts that they are unaware of.	Educate advisors and advisees about the unconscious nature of bias so they can detect conflicts of interest.
	Those being advised use this knowledge to make better decisions.	Advisees may view disclosures as indicating the advisor's objectivity.	Require disclosures for major sources of bias other than large financial ones.
	The duty to disclose also may motivate advisors to avoid conflicts of interest.	Advisors may act more biased in order to moderate the impact they expect their disclosures to have.	Have disclosures originate from a credible source other than the advisor.
Encouraging perspective-taking	People attempt to see the situation through the eyes of their adversaries.	Perspective-taking can lead people to a focus on the possible biases (e.g., self-interest) of their adversaries.	Use more direct perspective-taking techniques such as visualizing things through the other side's eyes.
	This leads them to understand their adversaries' perspective.	As a result, it can lead people to act more self-interestedly in order to counter the other side's self-interest.	Consider what valid reasons might underlie the other side's perspective.
	Fairer judgments result.	—	Consider how *you* would respond if *you* supported the other side.
Demanding objectivity	People are directly asked to be objective.	People are blind to their bias and thus assume they already are objective.	Educate people about the unconscious nature of bias so that they do not assume they would be aware if they were biased.
	This leads them to abandon or to correct for their biases.	Demands to focus on objectivity can increase people's confidence in their objectivity without curing their bias.	Encourage people to try to prevent bias before it occurs rather than to try to detect its occurrence after the fact.
	More impartial judgment results.	That confidence may license people to act yet more biased.	—

of awareness of their own biases. Consider the case discussed earlier of physicians and their potential to be biased by interactions with pharmaceutical companies in their patient-care decisions. One problem with mandatory disclosure is that because bias operates nonconsciously, physicians are unlikely to recognize potential conflicts of interest since they usually do not feel that these conflicts occur in the first place. For example, a pleasant dinner with a lovely pharmaceutical representative might leave one feeling more predisposed to the representative's wares, but if no cash changes hands, one might not recognize this as a conflict of interest. To the extent that clients are led to expect that the absence of conflict-of-interest disclosures implies the absence of bias, such lapses in disclosure could actually be worse than if disclosures were not required. They could provide medical patients (or other clients) with unwarranted confidence in their physicians' (or other advisors') objectivity when, in reality, that objectivity has been compromised by biases that are unrecognized and therefore undisclosed.

In cases where the requirement to disclose is in fact heeded, another potential hazard is likely to arise. That is, recipients of those disclosures may view them as a signal of the disclosers' objectivity. If one's stockbroker specifically mentions that she has a financial stake in the company she is recommending, one may

be particularly impressed by her objectivity in providing that disclosure. Even if one is aware that such a disclosure is required by law, one may nevertheless view it as signaling more about the stockbroker's integrity than about the requirements of the law. As discussed earlier, people display a powerful tendency to discount the importance of situational constraints in guiding others' behavior and instead tend to attribute that behavior to others' internal traits (Gilbert and Malone, 1995; Jones and Davis, 1965; Ross, 1977). Thus, the act of disclosing can have the perverse effect of making the discloser seem more trustworthy rather than more susceptible to bias (Cain, Loewenstein, and Moore, 2005).

Another problem with disclosure requirements is that they may have the ironic effect of making people's *behavior* more biased (Cain, Loewenstein, and Moore, 2005). Consider the case of an admissions officer evaluating an applicant who is the daughter of a close friend. Whereas he might think that the candidate should objectively be ranked in the top 20% of candidates, he might be inclined to instead rank her in the top 10% if he knows that he will have to disclose his potential for bias. That is, he might assume that others will discount the favorability of his view and that he should therefore offer a yet more favorable view in order to counterbalance that response. An experiment by Cain, Loewenstein, and Moore (2005) offers support for that hypothesis. Subjects were to estimate the monetary value of jars of coins and be rewarded for their accuracy. Before offering an estimate, each subject was able to quickly see the jars and also to avail him or herself of the advice of an "advisor" subject who had been given a far better look at the jars. The advisors were rewarded not for the estimators' accuracy but rather for how high the estimators' estimates were. The result was that the advisors suggested higher estimates (i.e., ones that were more biased toward their own self-interest) when they were required to disclose their conflict of interest. Moreover, their disclosures actually harmed those whom they advised, because those advisees actually made less profitable estimates than their peers, who also received biased advice but without such disclosures.

SUGGESTED SOLUTIONS

Research and theory regarding the bias blind spot and disclosure suggest that disclosure may backfire by inducing advisors to maintain their biases while paradoxically giving those whom they advise more confidence in their ethics and professionalism. Although solutions other than disclosure are probably necessary, some amendments to disclosure policies can help ensure that their benefits accrue to the individuals they aim to protect, rather than the experts who advise them.

Educate Disclosers

Disclosure policies should educate those subject to the requirement about the potential for unconscious bias in their judgments. Part of the problem is that disclosers typically are incapable of accurate self-assessment of their own biases (i.e., what must be disclosed). This applies to would-be disclosers including stockbrokers, real estate agents, judges, doctors, and scientific researchers. People need to understand that the requirement to disclose does not reflect the fact that people are aware of the factors that bias them, but rather that people often are not aware of the impact of those biasing factors. Because disclosers typically assume that bias involves an overt promotion of one's own self-interest, this alternative understanding must be taught for disclosers to understand the scope of information that is relevant to disclose.

Introduce Psychologically Savvy Disclosure Requirements

The belief that one is objective is so powerful that educating people about unconscious bias may not be enough to get disclosers to reveal all of the relevant information. Moreover, disclosers often have practical reasons for limiting the extent of their disclosure, such as avoiding bureaucratic hassles or the loss of valuable business (or, in a more beneficent case, a doctor worrying about being unable to persuade a patient to adopt the best therapy). To overcome these natural tendencies of disclosers, it is important to mandate the disclosure of specific information, not generic conflicts of interest. For example, doctors should be required to list the pharmaceutical companies from which they have received monetary compensation (and the approximate amounts involved), not just the relationships that might compromise their prescribing decisions. An added benefit of such a just-the-facts approach to gathering disclosure-relevant information is that it mitigates the tendency of the discloser to frame the information so as to avoid the appearance of conflict of interest (or, in an open-ended text disclosure, to obscure the relevant information in technical language).

Disclosure requirements need to include not only obvious cases of high stakes financial self-interest but also less obvious potential causes of bias. One example is the seemingly trivial gifts to doctors, such as pens with drug names on them. While doctors are indeed unlikely to be influenced by the economic value of the pen, psychological research suggests that repeatedly seeing the drug name is likely to unconsciously increase liking for, and prescribing of, the drug. Another

example is the potentially biasing effect of friendship. For example, Justice Scalia elected not to recuse himself from the case of *Cheney v. United States District Court for the District of Columbia* despite a close personal friendship with Vice President Cheney, claiming that the requirement for recusal did not apply in this case—that is, that his "impartiality [could not] reasonably be questioned." Without doubting the sincerity of Justice Scalia's belief in his own impartiality, the capacity to fully escape the unconscious biasing effects of friendship is a lot to expect.

As these examples illustrate, obtaining an accurate list of disclosures requires an understanding on the part of those who mandate them of the myriad of ways that bias can operate. These include not only the large sums of money that would be the key information from a standard economic perspective, but also social relationships, nonmonetary incentives, and personal beliefs not grounded in professional knowledge (e.g., religious or cultural values). Fortunately, psychologists have now itemized many of these factors, providing the requisite scientific knowledge for developing robust and effective disclosure requirements.

Educate Recipients

Disclosures can only be effective to the extent that those who receive them understand how bias operates and what effect it can have. Medical patients and financial investors who themselves rely on inaccurate theories of bias (e.g., Only big amounts of money could bias a well-paid doctor.) are unlikely to benefit from disclosures that rely on an accurate understanding of bias. Consequently, an important component of disclosures to clients (e.g., patients, investors) should involve not just a statement of the interest, but also a disclosure of the fact that interests of that sort have been found to bias advisors' guidance.

Use Third-Party Disclosers

Another problem with disclosures is that they typically are provided by the individual whose bias is in question. Thus, that individual benefits from the appearance of integrity and honesty afforded by that openness. For this reason, disclosures could be more beneficial if they came from a third-party source rather than from the potentially biased actor him or herself. Being told by a third party that one's physician has a financial stake in one's medical care could have a greater impact than being told that same thing by one's physician.

Consider, for example, the process of enrolling patients in clinical trials of new drugs. Often doctors receive a financial payment from the sponsoring pharmaceutical company for each patient that they enroll. Efforts (driven in large part by Senator Charles Grassley) are underway to mandate disclosure of all such payments in a central government database. For the disclosures to have maximum benefit, one would want to have the government (perhaps via email or automated phone calls or both) notify every patient considering participating in such trials of the financial benefit to their doctor. In contrast to disclosure by the doctor, which might make the doctor seem open and honest, the governmental notification could encourage the patient to reconsider the merits of the trial or to seek a second opinion. The disclosure is likely to be particularly helpful if it is accompanied by information about potential consequences of the relevant bias for the advisor's advice, as well as suggestions for alternative courses of action (such as seeking guidance from an independent physician).

Make Disclosures Readable and Transparent

One of the biggest problems with disclosures is that they are not transparent. Usually they are in tiny, almost unreadable print. Even if the type is readable, the meaning is often obscured by technical language or legalese. The underlying problems are twofold: first, disclosures can legitimately be complicated and extensive, and second, the disclosure statements are typically written by someone aiming to protect the discloser, not the client, resulting in the motivation to obfuscate (or to "overprotect" to the point of obscuring the most critical concerns). One approach to mitigating these problems is to mandate a short, nontechnical summary of the most important issues that is written not by the discloser but by a third-party clearinghouse or watchdog group. For understanding the full benefits and costs of complex products like mortgages or insurance policies, however, such a summary is bound to be insufficient. An alternative involves machine-readable disclosure (Thaler and Sunstein, 2008). The idea is that the disclosure is provided in a standardized computational form that allows third-party websites to compete so as best to reveal the key buried information. For example, if a consumer were choosing between two mortgage brokers, the computer could highlight only the relevant differences, greatly facilitating the client's ability to make an informed selection. Such standardized disclosures also deal with the discloser's lack of motivation to be transparent.

Accept that Disclosure Often Will Not Be Enough

Disclosures should generally be viewed as a first step, not a complete solution. Sometimes disclosure is necessary for identifying a conflict of interest but does nothing to remedy it. For example, judges ultimately make *decisions*, and when their reasoning

is susceptible to bias, mere disclosure provides no assurance that bias will not impact their decisions. Accordingly, the appropriate solution is recusal. For stockbrokers, while disclosure gives the client a chance to avoid taking actions driven by the advisor's bias, stronger precautions are typically needed in practice. For example, when a stockbroker has a stake in selling a particular stock, it may be acceptable for the broker to recommend the stock along with disclosure of the bias. However, prudent rules might prevent that broker from actually processing the client's transaction, which should instead be handled by another broker with no relationship to this one and no conflict of interest. Such a safeguard would prevent the client from being driven into a poor decision by social pressure. For doctors, bringing in a third party may also be useful. This is already done to a limited extent, for example, when a pharmacist recommends the generic form of a branded medication prescribed by a doctor. Having pharmacists phone or email doctors to check whether a patient could equally benefit from the least expensive of a set of comparable medications could also be beneficial. For more major medical decisions where the potential for bias exists, the best remedy is encouraging patients, as part of the disclosure, to obtain a second opinion.

Perspective Taking

The bias blind spot contributes to a host of important problems in the policy arena. Separate from the problems that it causes for disclosure (due to advisors' failure to recognize their own unconscious biases), another problem involves its propensity to exacerbate conflict and act as a barrier to effective negotiation. When people are convinced that objectivity is on their side, they are likely to resist compromising with those who disagree and instead to prefer a more aggressive response. One obvious solution, then, is to encourage people to consider others' perspectives. Trying to understand the opposing side's point of view, or to imagine the perspective one would take in that position, seems promising as a way to reduce the impact of the bias blind spot. For example, if Israelis could effectively understand the perspective of Palestinians, and vice versa, this could mitigate each side's propensity to see themselves as uniquely victimized and as having a monopoly on what constitutes an objectively fair resolution to their plight. Without that perspective, each side instead may feel justified in acting violently against a foe whom they view as too unreasonable to be negotiated with. Another example involves negotiations over legislation that impacts multiple interest groups (e.g., health-care reform, zoning laws). If each group feels that its position is objectively correct, each group will be ill-disposed to making compromises. Consideration of the other groups' perspectives may open up new avenues to reaching agreement, including "win-win" arrangements that effectively meet multiple parties' needs.

THE PROBLEM WITH PERSPECTIVE TAKING

Successful perspective taking has been shown to have a variety of positive effects relevant to conflict. It can increase people's altruism toward others, improve relationship satisfaction, decrease stereotypes about other groups, reduce self-serving judgments about what is fair, and produce more effective negotiation outcomes (e.g., Coke, Batson, and McDavis, 1978; Epley, Caruso, and Bazerman, 2006; Franzoi, Davis, and Young, 1985; Galinsky and Moskowitz, 2000; Neale and Bazerman, 1983; Savitsky et al., 2005). Unfortunately, solutions aimed at perspective taking can be difficult to implement successfully. For various psychological reasons, inducing people to successfully take others' perspectives is easier said than done. When the perspective taking is done poorly, its consequences can be worse than if efforts to do it were absent. A primary cause of this problem derives from individuals' naive realism, whereby they have difficulty separating their own subjective perceptions from what is true in objective reality. That tendency, combined with *biased assimilation* (people's inclination to carefully scrutinize, and ultimately reject, information that contradicts their prior beliefs) and self-serving biases (people's inclination to protect their ego even at the expense of accuracy), can make the process of thinking about adversaries' viewpoints result in people's becoming yet more convinced of the rightness of their own views (and the wrongness of their adversaries' views). Even worse, as a result of having made the effort to consider an adversary's perspective, one is likely to feel all the more righteous in championing one's own position. Consider the example of a conflict over land, such as that between the Israelis and Palestinians or between local groups involved in a zoning dispute. As each group makes the effort to consider the other side's perspective, they may be struck by the lack of "good reasons" for that side's position. As a consequence, their perspective-taking effort may lead them to feel even more strongly about their own position (i.e., to become *more* biased toward their own side)—even while their willingness to take the others' perspective makes them yet more convinced of their own objectivity.

There is another, related, potential downfall of perspective taking. People's efforts to consider another party's perspective can lead them to focus on and thereby exaggerate that party's biases and self-interest

(Epley, Caruso, and Bazerman, 2006). Consider the case of a contract negotiation between labor and management. If the manager attempts to take the perspective of the labor leader before sitting down for the negotiation, that manager might imagine that the labor leader will only think about making more money for the union workers and not at all about the financial needs of the firm. As a consequence, the manager's perspective-taking may make him yet more inclined to take a hard-line stance against wage increases out of concern that such a stance will be necessary in order to reach any reasonable middle ground. Thus, perspective-taking could lead one to take action that is more biased toward one's own side, if that perspective-taking leads one to believe that such selfish action is needed in order to counteract the bias of the other side and to thereby achieve an equitable outcome. Studies by Epley, Caruso, and Bazerman (2006) support this notion. In their studies, subjects were told to imagine the perspectives of different parties with interests contrary to their own or to consider their own perspective in a multiparty conflict over the allocation of scarce resources. The perspective-taking exercise was successful in that those who engaged in it thought it was fair for them to take less of a limited resource than did those who focused on their own perspective. Importantly, though, and inconsistent with those judgments, the subjects who engaged in perspective-taking *behaved* in a manner that was more biased than their peers: they took a larger portion of the scarce resource.

SUGGESTED SOLUTIONS

Simple instructions to consider the perspective of the other side can elicit perverse effects. However, more psychologically savvy efforts to encourage perspective taking can elicit desirable effects. A number of possible approaches are discussed below.

Encourage Cooperative Norms

Perspective-taking instructions are likely to be more effective when they are accompanied by a norm of cooperativeness rather than competitiveness (Epley, Caruso, and Bazerman, 2006). Because perspective taking changes people's opinions about what is fair in a direction that departs from their self-interest, having a norm that encourages fair behavior (rather than looking out for oneself) can capitalize on the benefits of perspective taking. A key challenge is how to institute this norm. Because adversaries are likely to enter a negotiation with differing motives and prior beliefs, it will typically be up to a third-party mediator to institute the cooperative norm. One possibility is for the mediator to provide incentives for cooperation. For example, in a contract negotiation, the two parties might agree in advance to an approach known as final-offer arbitration (which is used in major-league baseball negotiations). In it, each party proposes a solution, and the third-party arbitrator selects the more cooperative of the competing proposals, whose terms are then final. Another strategy for promoting cooperative norms involves framing by the third-party mediator. For example, in bringing together groups of influential Israelis and Palestinians for the purpose of working on conflict resolution, Herbert Kelman (e.g., Rouhana and Kelman, 1994) framed their task not as "negotiation" to resolve a "conflict," but rather as "joint problem solving." By using the "problem solving" frame, Kelman subtly introduced a norm that the parties' task was to cooperate rather than to aggressively advocate for their own side.

Manipulate Visual Perspective

The simplest perspective-taking instruction for third-party mediators to give is to tell adversaries to take each other's perspective. Because this simple instruction can backfire, it often is up to third parties to offer more nuanced perspective-taking instructions. For example, mediators might induce successful perspective taking by having adversaries imagine the visual perspective of those on the other side of the table from them, or even by having them take on that visual perspective by showing them a videotape of how that side is seeing the negotiation. This strategy involves inducing people to literally see the world from the other's perspective. Such manipulations of visual perspective have been shown to be powerful in changing people's judgments of others' personalities and behavior (e.g., Storms, 1973; Taylor and Fiske, 1975), and it is likely that those effects would extend to the domain of negotiation.

Use Carefully Worded Instructions

Effective perspective-taking instructions might induce people to imagine not how the other side sees the world, but rather what valid reasons there might be for them to see it that way (e.g., Puccio, 2003). As with the above strategy involving visual perspective taking, this method has the promise of eliciting perspective taking without immediately prompting adversaries to focus on the other side's bias. Because perspective-taking instructions can lead people to focus on their adversaries' bias, another solution is to lead people to instead focus on their adversaries' potential for objectivity. Since those on each side are likely to have preconceived notions about the other's bias with respect to the particular issue of their dispute, it may be more feasible to alter people's perceptions of the general objectivity of their adversary

Table 11.3 Financial benefits of seeing one's negotiation adversary as capable of objectivity rather than as biased

	"Objective" adversary	"Biased" adversary
Initial wage offer	$10.19	$10.08
Days of negotiating	9 days	15 days
Financial expense	$4.6 million	$7.0 million

Note: Differences between conditions were significant at the $p < .01$ level. Differences in management's initial wage offer mediated the effect of experimental condition on the financial expenses management incurred.

as a person, rather than to take aim directly at people's perceptions of their adversary's objectivity with respect to the conflict at hand. Kugler and Pronin (2007) tested that strategy. Subjects engaged in a wage contract negotiation where they represented management and their partner (unbeknown to them, actually a computer algorithm) represented labor. They were told that they would not be meeting their partner because the study concerned negotiation at a distance. Before beginning their negotiation, subjects were shown the results of a personality test allegedly taken by their adversary and designed to induce them to perceive that adversary either as prone to thinking about things objectively or as prone to bias (where the bias manipulation simply induced the same bias perceptions that normally arise in adversarial contexts). Subjects then began their negotiation by initiating the first round of bargaining. Bargaining continued until a wage agreement was reached, with substantial strike costs accruing to management for every "day" (i.e., round) without an agreement. The benefit to subjects of perceiving objectivity in their adversary was high (table 11.3). Had they been negotiating with real money, it would have amounted to $2.4 million. How did this happen? Subjects who believed in their adversary's potential for objectivity opened the negotiation with a fairer offer (their peers started with a lowball, highly competitive stance). As a result, they reached an agreement more quickly.

Limit Counterarguing

When adversaries present their perspectives to each other, they often have difficulty truly hearing what the other is saying. The reason is that people generally listen to their adversaries in a way that involves actively counterarguing those adversaries' reasoning rather than listening with an open ear and mind (e.g., Kunda, 1990; Lord, Ross, and Lepper, 1979). Such *counterarguing listening* involves activities such as judging the problems and weaknesses in the other's position as he or she is stating it, thinking about ways in which one's own position is superior, and preparing counterarguments that can be leveled when it is one's chance to reply. While third-party mediators typically are motivated to listen with an open ear and mind, adversaries typically are motivated to devote their mental and verbal energy to discrediting or countering the points made by their adversary, even though this strategy prevents effectively hearing the other's perspective. Thus, third-party mediators could encourage better perspective taking by inducing adversaries to listen without counterarguing. In a recent experiment, Kennedy and Pronin (2009) aimed to elicit such listening with a simple instruction. Participants were faced with a fellow student who held a different position from their own on a campus issue (one involving academic grading practices). Those in the condition designed to reduce counterarguing listening were told that after hearing from their adversary, they would be asked to accurately repeat, in their own words, the details of their adversary's position on the issue—such that their adversary would agree that his position was "accurately captured and represented." The experiment revealed that those faced with this task came to view their adversary as less biased, and more objective, than did those left to counterargue their adversary as they listened.

Encourage People to "Consider the Opposite"

A different strategy for inducing more effective perspective-taking has been called the *consider the opposite* strategy (Lord, Lepper, and Preston, 1984). That strategy does not ask people to consider events from the other side's perspective but rather it induces them to do so. One experiment testing that strategy was conducted in the context of the polarizing issue of the death penalty. The experiment was inspired by earlier research showing that people exposed to mixed evidence about the effectiveness of the death penalty as a deterrent generally come to feel yet more strongly for their own side (Lord, Ross, and Lepper, 1979). In order to induce people to take a less biased and more evenhanded view of such evidence, the researchers sought to induce them to take the opposing perspective on that evidence. However, rather than directly asking them to look at things from the other side's point of view, they instead asked them to read each piece of evidence and "ask yourself at each step whether you would have made the same high or low evaluations had exactly the same study produced results on the other side of the issue" (p. 1233). With this instruction, the subjects no longer showed the usual bias toward their own side. In a simple control condition where they received no instructions, and in a comparison condition where they were instructed to be *objective* and *unbiased*, they instead showed the usual bias effect.

Demanding Objectivity

Instructions to be objective and unbiased would seem to be a straightforward way to encourage that behavior. They do not rely on individuals' ability to perspective take, and they attempt to remove people's biases rather than to simply have people disclose them. However, research studies (including the one described in the preceding paragraph) have shown that simple pleas for objectivity do not work and can even induce perverse effects (Frantz and Janoff-Bulman, 2000; Lord, Lepper, and Preston, 1984; Wilson et al., 1996).

THE PROBLEM WITH DEMANDING OBJECTIVITY

The problem with demanding objectivity rests on the unconscious nature of bias. Because people are not typically aware of their biases, they are not in a position to respond to instructions to consciously eliminate those biases. Indeed, such instructions may instead have the opposite effect of causing people to be more biased; that is, individuals are likely to respond to those instructions by looking inward for signs of bias and, upon finding none, feeling yet more confident in their own objectivity. That confidence is apt to make them become more biased by preventing them from feeling the need to engage in the sort of questioning and examination that might help them understand the views of the other side.

In a series of studies by Uhlmann and Cohen (2007), subjects primed to feel personally objective (by completing a scale in which they were able to assert their characterological objectivity) were more likely to show gender-biased discrimination in the context of a hypothetical hiring decision. A series of studies by Frantz and Janoff-Bulman (2000) also support the hypothesis that feelings of objectivity do not guarantee actual objectivity and in some cases can be indicative of increased bias. In those studies, subjects read various conflict scenarios that manipulated the likeability of those on opposing sides of a conflict. To the extent that participants liked one of the individuals more than the other, they tended to claim that the individual whom they liked was on the right side of the conflict. Importantly, instructions to be objective only exacerbated this bias. Apparently, the subjects had an automatic (and nonconscious) tendency to view the likeable person in the scenario as the objectively correct one. As a result, the instruction to be objective only led them to feel more strongly in favor of the side that they viewed as objectively correct. Taken together, these two sets of studies illustrate that in both social conflict and employment-discrimination settings, people view their biased perspectives as objective and become more biased as their confidence in their own objectivity is raised.

Despite these findings, instructions to be objective continue to be common. For example, in the legal arena, judges typically provide instructions to jurors such as: "Do not allow sympathy or prejudice to influence you. The law demands of you a just verdict, unaffected by anything except the evidence, your common sense, and the law as I give it to you" (U.S. District Court, 8th Circuit, 2007). Such instructions are liable to augment, rather than mitigate, the impact of juror biases on overall jury decisions. Judges, in turn, in deciding whether to recuse themselves, are required to ask themselves whether they can be objective with respect to a particular case; that is, whether their impartiality could reasonably be questioned. When one's answer is in the negative, despite the presence of unconscious bias, the very process of deciding that one is objective enough to hear a case may tend to magnify one's bias. In light of these concerns, Judge Richard Posner of the U.S. Court of Appeals (2008) argued against the wisdom of criteria that rely on judges' ability to internally assess their own bias.

SUGGESTED SOLUTIONS

The problem with encouraging people to be objective is that they generally already take for granted that they are being just that. Accordingly, the first solution discussed below involves educating people about the unconscious nature of bias. Other solutions can be used in conjunction with such education, or on their own, in order to lead people to exhibit increased objectivity.

Educate about Unconscious Bias

A starting point is to teach people that bias typically operates outside of conscious awareness. Doing so can help people to recognize their susceptibility to bias by preventing them from relying excessively on introspective evidence of bias. Furthermore, it can reduce the bias blind spot by helping people realize that they are not likely to be any less biased than those around them. It also can inspire people to engage in efforts to overcome their biases. Research by Pronin and Kugler (2007) has suggested the promise of this strategy. In one experiment, subjects either read an article informing them about the role of nonconscious processes in judgment and about people's lack of awareness of being influenced by those processes, or they were in a control condition in which they did not read that article (both groups also read a filler article masking the researchers' true interests). Then, in an allegedly separate experiment, participants were asked to indicate their personal susceptibility relative to

their student peers to a variety of different judgmental biases. The result was that participants who had been educated about nonconscious processes (and about the perils of relying on introspection) saw themselves as no more objective than their peers, unlike those in the control condition. The two conditions differed significantly from each other, indicating that the intervention reduced the bias blind spot.

Reduce Exposure to Biasing Information

Given that bias typically operates nonconsciously, it is preferable to avoid exposure to biasing information rather than to try to correct for such exposure after the fact. For example, it would be next to impossible for a teacher to grade the papers of a very nice and not-so-nice student objectively, without over- or undercorrecting for the impact of the student's niceness. It would be straightforward, however, to grade the papers blindly, thereby removing the risk of bias. Similarly, when watching an orchestra musician play on stage, it might be difficult to judge his or her musicality without being biased by appearance and gender. The now widely used practice of having such musicians audition behind a curtain successfully removes this risk of bias (and, not incidentally, has led to dramatic advances for female orchestra players). Similar logic underlies the FDA's requirement for double blind methods (i.e., for both health-care professionals and their patients) in the clinical trials required for drug approval. When we choose to grade papers blindly, to judge musicians blindly, or to conduct clinical trials blindly, we do so not because we can feel our expectations biasing our grading, or our gender stereotypes biasing our judgments of musicality, or our desires for drug approval biasing our clinical evaluations, but rather because we recognize that the lack of those feelings does not necessarily signal a lack of bias.

Demand Behavior That Would Appear Objective to an Outsider

To the extent that exposure to biasing information cannot be avoided, a modified form of the standard demand to be objective has merit. That modification involves asking people not to be assured of their own objectivity, but rather to be assured that *others* will see them as objective. Thus, the instruction could be something like: strive to make your behavior look objective to an outside observer. Such a strategy is used, for example, when individuals coaching people who are dealing with ethical dilemmas advise them to ask themselves whether they would be happy with their decision being reported on the front page of the newspaper. This instruction is intended to lead people to evaluate the ethicality of their decisions not by looking inward to determine whether they have been biased by self-interest, but by looking outward to determine whether others would have that opinion. The difference between striving to be objective versus striving to be viewed as objective by an outsider is a key one, because the former involves assessing the presence of bias by looking inward to conscious thoughts and motives, whereas the latter involves looking to observable actions to make that determination of bias. Due to the unconscious nature of bias, strategies that involve looking inward are likely to miss bias when it is present, whereas strategies that involve looking to outward behavior are more likely catch it. Finally, the impact of this sort of objectivity instruction could be further enhanced by reminding people that this instruction is not as strange as it might initially sound—since while one may be inclined to judge one's own objectivity based on what's in one's head, the rest of the world will judge it by looking at one's actions.

Concluding Thoughts

Over the past several decades, psychologists have documented a wide range of biases that influence people's thoughts, judgments, and behavior. In addition to the problems that these biases can cause, more recent evidence has highlighted the problems associated with people's biased perceptions of their own (and others') biases. People show a bias toward recognizing bias more in those around them than in themselves. This bias blind spot can elicit and exacerbate a range of policy-relevant problems. Policies that target problems in domains varying from ethical lapses to discrimination to conflict can be informed by knowledge about this asymmetry in people's perceptions of bias.

The human mind is unlikely to free itself of the biases that take hold of it. Indeed, those biases can sometimes serve valuable functions such as allowing people to maintain healthy self-esteem and to form judgments quickly with a minimal expenditure of mental resources. And, to the extent that these biases are beneficial, it may be just as well that individuals maintain a blissful lack of awareness of their commissions of them. However, this lack of awareness becomes a problem when individuals would be better off correcting for or warding off their biases, and when individuals impute bias to others that they deny in themselves. At a collective level, people's shared blindness to their biases can exert particularly damaging effects because entire institutions can succumb to biases of which each individual contributor is unaware. In such cases, individuals might benefit from recognizing that their own minds are unlikely to be free of the biases that they so readily observe taking hold of the minds of those around them.

Acknowledgment

This research was supported by grants to E. Pronin from the National Science Foundation (BCS-0742394) and from the FINRA Investor Education Foundation.

References

Agrawal, S., Saluja, I., and Kaczorowski, J. (2004). A prospective before-and-after trial of an educational intervention about pharmaceutical marketing. *Academic Medicine*, *79*, 1046–1050.

Alicke, M. D., and Govorun, O. (2005). The better-than-average effect. In M. D. Alicke, D. A. Dunning, and J. I. Krueger (Eds.), *The self in social judgment* (pp. 85–106). New York: Psychology Press.

Armor, D. A. (1999). The illusion of objectivity: A bias in the perception of freedom from bias. *Dissertation Abstracts International*, *59*, 5163B.

Bazerman, M. H. (2002). *Judgment in managerial decision making* (5th ed.). New York: Wiley.

Bazerman, M. H., Loewenstein, G., and Moore, D. A. (2002, November). Why good accountants do bad audits. *Harvard Business Review*, pp. 96–102.

Bornstein, R. F. (1989). Exposure and affect: Overview and meta-analysis of research, 1968–1987. *Psychological Bulletin*, *106*, 265–289.

Buehler, R., Griffin, D., and Ross, M. (1994). Exploring the "planning fallacy": Why people underestimate their task completion times. *Journal of Personality and Social Psychology*, *67*, 366–381.

Cain, D. M., Loewenstein, G., and Moore, D. A. (2005). The dirt on coming clean: Perverse effects of disclosing conflicts of interest. *Journal of Legal Studies*, *34*, 1–25.

Caruso, E., Epley, N., and Bazerman, M. H. (2006). The costs and benefits of undoing egocentric responsibility assessments in groups. *Journal of Personality and Social Psychology*, *91*, 857–871.

Cohen, G. L. (2003). Party over policy: The dominating impact of group influence on political beliefs. *Journal of Personality and Social Psychology*, *85*, 808–822.

Coke, J. S., Batson, C. D., and McDavis, K. (1978). Empathic mediation of helping: A two-stage model. *Journal of Personality and Social Psychology*, *36*, 752–765.

Crenshaw, M. (1998). The logic of terrorism: Terrorist behavior as a product of strategic choice. In W. Reich (Ed.), *Origins of terrorism: Psychologies, ideologies, theologies, states of mind* (pp. 7–24). Princeton, NJ: Woodrow Wilson Center Press.

Dana, J., and Loewenstein, G. (2003). A social science perspective on gifts to physicians from industry. *Journal of the American Medical Association*, *290*, 252–255.

Dawson, E., Gilovich, T., Regan, D. T. (2002). Motivated reasoning and the Wason selection task. *Personality and Social Psychology Bulletin*, *28*, 1379–1387.

Diaz-Sprague, R. (2003). The MIT success story: Interview with Nancy Hopkins. *AWIS Magazine*, *32*, 10–15.

Ditto, P. H., and Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions, *Journal of Personality and Social Psychology*, *62*, 568–584.

Dovidio, J. F., and Gaertner, S. L. (1991). Changes in the expression and assessment of racial prejudice. In H. J. Knopke, R. J. Norrell, and R. W. Rogers (Eds.), *Opening doors: Perspectives of race relations in contemporary America* (pp. 119–148). Tuscaloosa: University of Alabama Press.

———. (2004). Aversive racism. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 36, pp. 1–52). San Diego, CA: Academic Press.

Dunning, D., Griffin, D. W., Milojkovic, J. D., and Ross, L. (1990). The overconfidence effect in social prediction. *Journal of Personality and Social Psychology*, *58*, 568–581.

Dunning, D., Johnson, K., Ehrlinger, J., and Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, *12*, 83–87.

Dunning, D., Meyerowitz, J. A., and Holzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving appraisals of ability. *Journal of Personality and Social Psychology*, *57*, 1082–1090.

Ehrlinger, J., Gilovich, T., and Ross, L. (2005). Peering into the bias blind spot: People's assessments of bias in themselves and others. *Personality and Social Psychology Bulletin*, *31*, 680–692.

Epley, N., Caruso, E. M. and Bazerman, M. H. (2006). When perspective taking increases taking: Reactive egoism in social interaction. *Journal of Personality and Social Psychology*, *91*, 872–889.

Epley, N., and Dunning, D. (2000). Feeling "holier than thou": Are self-serving assessments produced by errors in self- or social prediction? *Journal of Personality and Social Psychology*, *79*, 861–875.

Fazio, R. H., and Olsen, M. A. (2003). Implicit measures in social cognition: Their meaning and use. *Annual Review of Psychology*, *54*, 297–327.

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, *1*, 288–299.

Frantz, C. P. (2006). I AM being fair: The bias blind spot as a stumbling block to seeing both sides. *Basic and Applied Social Psychology*, *28*, 157–167.

Frantz, C. M., and Janoff-Bulman, R. (2000). Considering both sides: The limits of perspective taking. *Basic and Applied Social Psychology*, *22*, 31–42.

Franzoi, S. L., Davis, M. H., and Young, R. D. (1985). The effect of private self-consciousness and perspective-taking on satisfaction in close relationships. *Journal of Personality and Social Psychology*, *48*, 1584–1594.

Friedrich, J. (1996). On seeing oneself as less self-serving than others: The ultimate self-serving bias? *Teaching of Psychology*, *23*, 107–109.

Galinsky, A. D., and Moskowitz, G. B. (2000). Perspective taking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of Personality and Social Psychology*, *78*, 708–724.

Gilbert, D. T., and Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin*, *117*, 21–38.

Goethals, G. R. (1986). Fabricating and ignoring social reality: Self-serving estimates of consensus. In J. M. Olsen, C. P. Herman, and M. P. Zanna (Eds.), *Relative deprivation and social comparison: The Ontario Symposium* (Vol. 4, pp.135–157). Hillsdale, NJ: Erlbaum.

Gorn, G. J., Goldberg, M. E., and Basu, K. (1993). Mood, awareness, and product evaluation. *Journal of Consumer Psychology*, *2*, 237–256.

Greenwald, A. G., and Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*, 4–27.

Heath, C. (1999). On the social psychology of agency relationships: Lay theories of motivation overemphasize extrinsic incentives. *Organizational Behavior and Human Decision Processes*, *78*, 25–62.

Jones, E. E., and Davis, K. E. (1965). From acts to dispositions: The attribution process in person perception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 2, pp. 219–266). New York: Academic.

Jones, E. E., and Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, *3*(1), 1–24.

Jones, E. E., and Nisbett, R. E. (1972). The actor and the observer: Divergent perceptions of the cause of behavior. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, and B. Weiner (Eds.), *Attribution: Perceiving causes of behavior* (pp. 79–94). Morristown, NJ: General Learning Press.

Kahneman, D., and Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*, 237–251.

———. (1979). Prospect theory: An analysis of decisions under risk. *Econometrica*, *47*, 313–327.

———. (1982). On the study of statistical intuitions. *Cognition*, *11*, 123–141.

Katz, D., Mansfield, P., Goodman, R., Tiefer, L., and Merz, J. (2003). Psychological aspects of gifts from drug companies. *Journal of the American Medical Association*, *290*, 2404.

Kennedy, K. A., and Pronin, E. (2007). [Disagreement with suicide bombers' goals and concerns.] Unpublished data.

———. (2008). When disagreement gets ugly: Perceptions of bias and the escalation of conflict. *Personality and Social Psychology Bulletin*, *34*, 833–848.

———. (2009). [Non-counterarguing listening and bias perception.] Unpublished data.

———. (2012). Bias perception and the spiral of conflict. In J. Hanson (Ed.), *Ideology, psychology, and law* (pp. 410–466). Oxford: Oxford University Press.

Krueger, J. (1998). Enhancement bias in description of self and others. *Personality and Social Psychology Bulletin*, *24*, 505–516.

Kruger, J., and Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*, 1121–1134.

Kruger, J., and Gilovich, T. (1999). Naïve cynicism in everyday theories of responsibility assessment: On biased assumptions of bias. *Journal of Personality and Social Psychology*, *76*, 743–753.

Kugler, M. B., and Pronin, E. (2007). *The benefits of perceiving objectivity in a negotiation adversary.* Manuscript in preparation.

Kunda, Z. (1987). Motivated inference: Self-serving generation and evaluation of causal theories. *Journal of Personality and Social Psychology*, *53*, 636–647.

———. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. doi:10.1037/0033-2909.108.3.480

Kwan, V.S.Y., John, O. P, Kenny, D. A., Bond, M. H., and Robins, R. W. (2004). Reconceptualizing individual differences in the self-enhancement bias: An interpersonal approach. *Psychological Review*, *111*, 94–110.

Leo, J. (2002, April 8). Bogus bias at MIT. *U.S. News and World Report*, p. 43.

Lord, C. G., Lepper. M. R., and Preston, E. (1984). Considering the opposite: A corrective strategy, *Journal of Personality and Social Psychology*, *47*, 1231–1243.

Lord, C. G., Ross, L., and 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.

Margalit, A. (2003, January 16). The suicide bombers. *New York Review of Books*, p. 50.

Massachusetts Institute of Technology, Committee on Women Faculty in the School of Science (1999). A study on the status of women faculty in science at MIT. *MIT Faculty Newsletter*, *11* (Special Ed).

Mayer, J. D., Gaschke, Y., Braverman, D, L., and Evans, T. (1992). Mood-congruent judgment is a general effect. *Journal of Personality and Social Psychology*, *63*, 119–132.

McKinney, W., Schiedermayer, D., Lurie, N., Simpson, D., Goodman, J. and Rich, E. (1990). Attitudes of internal medicine faculty and residents toward professional interaction with pharmaceutical sales representatives.

Journal of the American Medical Association, *264*, 1693–1697.

Merari, A. (2004). *Suicide terrorism in the context of the Israeli-Palestinian conflict.* Paper commissioned for Suicide Terrorism Conference. Washington, DC: National Institute of Justice.

Miller, A. G., Baer, R., and Schonberg, P. (1979). The bias phenomenon in attitude attribution: Actor and observer perspectives. *Journal of Personality and Social Psychology*, *37*, 1421–1431.

Miller, D. T. (1999) The norm of self-interest. *American Psychologist*, *54*, 1053–1060.

Miller, D. T., and Ratner, R. K. (1998). The disparity between the actual and assumed power of self-interest. *Journal of Personality and Social Psychology*, *74*, 53–62.

Miller, D. W., and Wilson, R. (1999). MIT acknowledges bias against female faculty members. *Chronicle of Higher Education*, *45*, A18.

Moore, D. A., Tetlock, P. E., Tanlu, L., and Bazerman, M. H. (2006). Conflicts of interest and the case of auditor independence: Moral seduction and strategic issue cycling. *Academy of Management Review*, *31*, 10–29.

Morgan, M. A., Dana, J., Loewenstein, G. M., Zinberf, S., and Schulkin, J. (2006). Interactions of doctors with the pharmaceutical industry. *Journal of Medical Ethics*, *32*, 559–563.

Neale, M. A., and Bazerman, M. H. (1983). The effects of perspective taking ability under alternative forms of arbitration on the negotiation process. *Industrial and Labor Relations Review*, *26*, 378–388.

Nisbett, R. E., Borgida, U., Crandall, R., and Reed, H. (1976). Popular induction: Information is not necessarily informative. In J. W. Payne, and J. S. Carroll (Eds.), *Cognition and social behavior* (pp. 113–133). Hillsdale, NJ: Erlbaum Associates.

Nisbett, R. E., and Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*, 231–259.

Norton, M. I., Vandello, J. A., and Darley, J. M. (2004) Casuistry and social category bias. *Journal of Personality and Social Psychology*, *87*, 817–831.

Paese, P. W., and Yonker, R. D. (2001). Toward a better understanding of egocentric fairness judgments in negotiation. *International Journal of Conflict Management*, *12*, 114–131.

Pape, R. (2005). *Dying to win: The strategic logic of suicide terrorism.* New York: Random House.

Posner, R. A. (2008). *How judges think*. Cambridge, MA: Harvard University Press.

Post, J. (2005). Psychological operations and counterterrorism. *Joint Forces Quarterly*, *37*, 105–110.

Pronin, E. (2007). Perception and misperception of bias in human judgment. *Trends in Cognitive Sciences*, *11*, 37–43.

———. (2009). The introspection illusion. *Advances in experimental social psychology* (Vol. 41, pp. 1–67). San Diego, CA: Elsevier.

Pronin, E., Berger, J. A., and Molouki, S. (2007). Alone in a crowd of sheep: Asymmetric perceptions of conformity and their roots in an introspection illusion. *Journal of Personality and Social Psychology*, *92*, 585–595.

Pronin, E., Gilovich, T., and Ross, L. (2004). Objectivity in the eye of the beholder: Divergent perceptions of bias in self versus others. *Psychological Review*, *111*, 781–799.

Pronin, E., Gray, C., and Shepherd, H. (2007). *The group target effect in perceptions of personal bias.* Manuscript submitted for publication.

Pronin, E., Kennedy, K., and Butsch, S. (2006). Bombing versus negotiating: How preferences for combating terrorism are affected by perceived terrorist rationality. *Basic and Applied Social Psychology*, *28*, 385–392.

Pronin, E., and Kugler, M. B. (2007). Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot. *Journal of Experimental Social Psychology*, *43*, 565–578.

Pronin, E., Lin, D. Y., and Ross, L. (2002). The bias blind spot: Perception of bias in self versus others. *Personality and Social Psychology Bulletin*, *12*, 83–87.

Puccio, C. T. (2003). The search for common ground: Overcoming false polarization and unwarranted pessimism about ideological differences. *Dissertation Abstracts International*, *64*, 2444B.

Reeder, G. O., Pryor, J. B., Wohl, M.J.A., and Griswell, M. L. (2005). On attributing negative motives to others who disagree with our opinions. *Personality and Social Psychology Bulletin*, *31*, 1498–1510.

Reich, D. A. (2004). What you expect is not always what you get: The roles of extremity, optimism and pessimism in the behavioral confirmation process. *Journal of Experimental Social Psychology*, *40*, 199–215.

Robinson, R. J., Keltner, D., Ward, A., and Ross, L. (1995). Actual versus assumed differences in construal: "Naïve realism" in intergroup perception and conflict. *Journal of Personality and Social Psychology*, *68*, 404–417.

Roese, N. J., and Olson, J. M. (2007). Better, stronger, faster: Self-serving judgment, affect regulation, and the optimal vigilance hypothesis. *Perspectives on Psychological Science*, *2*, 124–141

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 10, pp. 173–220). New York: Academic Press.

Ross, L., McGuire, J., and Minson, J. (2004). *The relationship between self-other disagreement and the perceived impact of biasing versus normative consideration on own versus others' opinions.* Unpublished manuscript.

Ross, L., and Ward, A. (1995). Psychological barriers to dispute resolution. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 27, pp. 255–304). San Diego, CA: Academic Press.

Ross, M., and Sicoly, F. (1979). Egocentric biases in availability and attribution. *Journal of Personality and Social Psychology*, 37, 322–336.

Rouhana, N. N., and Kelman, H. C. (1994). Promoting joint thinking in international conflicts: An Israeli-Palestinian continuing workshop. *Journal of Social Issues*, *50*, 157–178.

Savitsky, K., Van Boven, L. Epley, N., and Wight, W. (2005). The unpacking effect in responsibility allocations for group tasks. *Journal of Experimental Social Psychology*, *41*, 447–457.

Sedikides, C. (2007). Self-enhancement and self-protection: Powerful, pancultural, and functional. *Hellenic Journal of Psychology*, *4*, 1–13.

Sherman, D. K., Nelson, L. D., and Ross, L. D. (2003). Naïve realism and affirmative action: Adversaries are more similar than they think. *Basic and Applied Social Psychology*, *25*, 275–289.

Snyder, M., and Swann, W. B., Jr. (1978). Hypothesis-testing processes in social interaction. *Journal of Personality and Social Psychology*, *36*, 1202–1212.

Son Hing, L. S., Chung-Yan, G. A., Grunfeld R., Robichaud, L., and Zanna, M. P. (2005). Exploring the discrepancy between implicit and explicit prejudice: A test of aversive racism theory. In J. P. Forgas, K. Williams, and S. Latham (Eds.), *Social motivation, conscious and unconscious processes* (pp. 275–293). New York: Psychology Press.

Son Hing, L. S., Li, W., and Zanna, M. P. (2002). Inducing hypocrisy to reduce prejudicial responses among aversive racists. *Journal of Experimental Social Psychology*, *38*, 71–78.

Sprinzak, E. (2000, September-October). Rational fanatics. *Foreign Policy*, pp. 66–73.

Storms, M. (1973). Videotape and the attribution process: Reversing actors' and observers' points of view. *Journal of Personality and Social Psychology*, 27, 165–175.

Tajfel, H., and Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin and S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33–47). Monterey, CA: Brooks/Cole.

Taylor, S. E., and Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, *103*, 193–210.

Taylor, S. E., and Fiske, S. T. (1975). Point of view and perceptions of causality, *Journal of Personality and Social Psychology*, *32*, 439–445.

Thaler, R. H, and Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness.* New Haven, CT: Yale University Press.

Thompson, L., and Nadler, J. (2000). Judgmental biases in conflict resolution and how to overcome them. In P. T. Coleman and M. Deutsch (Eds.), *The handbook of conflict resolution: Theory and practice* (pp. 213–235). San Francisco: Jossey-Bass.

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1130.

Uhlmann, E., and Cohen, G. L. (2007). "I think it, therefore it's true": Effects of self-perceived objectivity on hiring discrimination. *Organizational Behavior and Human Decision Processes*, *104*, 207–223.

———. (2005). Constructed criteria: Redefining merit to justify discrimination. *Psychological Science*, *16*, 474–480.

U.S. District Court, 8th Circuit. 8th Cir. Civil Jury Instr. § 4.50A (2007).

Van Boven, L., Kamada, A., and Gilovich, T. (1999). Their perceiver as perceived: Everyday intuitions about the correspondence bias. *Journal of Personality and Social Psychology*, *77*, 1188–1199.

Van Boven, L., White, K., Kamada, A., and Gilovich, T. (2003). Intuitions about situational correction in self and others. *Journal of Personality and Social Psychology*, *85*, 249–258.

Vivian, J. E., and Berkowitz, N. H. (1992). Anticipated bias from an outgroup: An attributional analysis. *European Journal of Social Psychology*, *22*, 415–424.

Wazana, A. (2000). Physicians and the pharmaceutical industry: Is a gift ever just a gift? *Journal of the American Medical Association*, *283*, 373–380.

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, *39*, 806–820.

Wilson, T. D., Centerbar, D. B., and Brekke, N. (2002). Mental contamination and the debiasing problem. In T. Gilovich, D. Griffin, and D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 185–200). New York: Cambridge University Press.

Wilson, T. D., and Gilbert, D. T. (2003). Affective forecasting. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 35, pp. 345–411). San Diego, CA: Academic Press.

Wilson, T. D., Houston, C. E., Etling, K. M., and Brekke, N. (1996). A new look at anchoring effects: Basic anchoring and its antecedents. *Journal of Experimental Psychology: General*, *125*, 387–402.

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, *9*, 1–2.

CHAPTER 12

Questions of Competence

The Duty to Inform and the Limits to Choice

BARUCH FISCHHOFF

SARA L. EGGERS

Many of our decisions are shaped by government policies that reflect policy makers' beliefs about our competence to make those choices. For example, policies establishing disclosure requirements for investments and pharmaceuticals reflect beliefs about our competence to recruit and comprehend the relevant evidence. Policies regulating the claims made about consumer products and political candidates reflect beliefs about our competence to evaluate them. Policies governing living wills reflect beliefs about our competence to anticipate personally unprecedented circumstances.

The stakes riding on these beliefs are high. If our competence is overestimated, then we may be denied needed protections. If our competence is underestimated, then we may be wrongly denied the right to choose. If ways to enhance our competence are underestimated, then we may lose chances for empowerment. If those opportunities are overestimated, then we may receive useless help (e.g., overly complex financial or medical disclosures) and be held responsible for its failure.

Judgments of decision-making competence are subject to known biases (Gilovich, Griffin, and Kahneman, 2002). *Outcome bias* leads to assessing decisions by the outcomes that follow them rather than by the thinking that goes into them. As a result, people facing easy choices (e.g., among places to eat) seem more competent than people facing hard ones (e.g., among medical treatments). *Hindsight bias* leads to exaggerating the competence of people who experience good fortune and underestimating that of those who do not. *Defensive attribution* leads to deprecating the competence of people whom misery befalls, so that observers can feel less vulnerable to suffering a similar fate.

Beliefs about decision-making competence can also reflect motivated thinking (or even deliberate misrepresentation), when a policy's legitimacy depends on the perceived competence of those whose behavior it governs. For example, advocates of market-based policies see consumers (investors, patients, etc.) as competent, as do advocates of participatory policies (e.g., referenda, citizen advisory committees). Advocates of strong regulatory policies see consumers (investors, patients, etc.) as incompetent and in need of that protection, as do advocates of policies that empower technocratic elites. Advocates of reproductive rights for adolescents make strong claims for teens' competence; opponents of adjudicating teens as adults make contrary claims (*Roper v. Simmons*, 2005). Both cannot be right, at least without more discriminating accounts of the similarities and differences in these decisions and the teens making them (Parker and Fischhoff, 2005).

Sweeping generalizations about decision-making competence can stimulate useful public discourse by encouraging partisans to assemble and defend examples supporting their positions. However, strategically chosen and interpreted examples are just that: things to consider, not systematic evidence. Sound policies require detailed analyses that are able to capture the heterogeneity in both the demands that decisions make and the skills of those making them (Bruine de Bruin, Parker, and Fischhoff, 2007).

An Approach to Competency-Based Policy Making

This chapter offers such a general approach to assessing and, where possible, improving, individuals' competence to make specific decisions under the conditions created by specific policies. It illustrates the approach with risk-related decisions in U.S. policy contexts that were chosen to suggest the variety of

possible incentives and opportunities for implementing policies that enhance public decision-making competence. The examples cover a variety of topics (including drugs, pathogens, and contaminants), their policy-making locus (including regulators, courts, and emergency officials), and their decision makers (including teens, older men, hobbyists, and everyone). They are presented in roughly decreasing order of how explicit the decisions are, beginning with ones made at a clear point in time and ending with ones embedded in the flow of events.

Our approach follows the "traditional" strategy of behavioral decision research (Edwards, 1954; von Winterfeldt and Edwards, 1986):

1. *Normative analysis*: identifying the best choices, using the available science to predict the outcomes of possible choices and decision makers' values to weight them
2. *Descriptive analysis*: predicting the choices that those individuals would actually make under the conditions created by possible policies
3. *Prescriptive analysis*: characterizing the gap between the normative ideal and the descriptive reality with each policy

Determining the prescriptive implications of a normative-descriptive gap requires a value judgment. Among other things, policy makers must weigh the fates of different individuals. They might treat everyone equally or assign weights based on properties like age, health, pregnancy status, citizenship, or historical injustices (Fischhoff, Atran, and Fischhoff, 2007). Consider, for example, a choice between policies requiring just English or both English and Spanish on warning labels. Holding the font size constant, the former allows more words, so that labels can address more problems or the same problems more thoroughly. However, that policy leaves Spanish-only speakers less protected. A third policy, reducing the font size, could accommodate both languages but lose users with limited vision or aversion to fine print. A fourth policy, expanding the size of the warning label, could allow more words or larger font but crowd out benefit information. The prescriptive analysis makes these choices more explicit—which policy makers may or may not welcome.

The approach also allows clarifying the impacts of policies that honor procedural principles, such as "freedom of choice," "consumer protection," "chances to learn from experience," or "full disclosure." For example, a First Amendment right to "commercial freedom of speech" has been invoked to expand the range of legal product claims. The result could be positive if consumers can interpret the claims, negative if not. "Full disclosure" has been advocated as a way to extract needed information from producers. The result could be positive if consumers can extract the decision-relevant facts, negative if the clutter overwhelms them. The approach characterizes such policies by their effects on consumers' ability to make the choices that they govern.

Applying the approach requires contributions from multiple disciplines. Identifying the optimal choice requires decision analysis informed by behavioral research (capturing individuals' values) and subject matter expertise (regarding expected outcomes). Predicting individuals' choices requires behavioral research into how individuals interpret the choices that emerge under different policies. Evaluating the gap between the normative ideal and the descriptive reality requires expertise in law and philosophy. Choosing policies requires political judgment informed by scientific assessments of impacts on outcomes that matter to policy makers.

Although described as sequential, these steps are inherently interdependent. Without knowing individuals' values, analysts cannot identify evidence relevant to their choices. Without knowing policy makers' values, analysts cannot properly disaggregate outcomes (e.g., by age, gender). Without knowing the sources of poor choices (e.g., lack of skills, facts, or motivation), policy makers cannot understand their options. And so on.

The following case studies illustrate the approach with risk-related policies. Other applications include avian flu (Fischhoff et al., 2006), sexual assault (Fischhoff, 1992), nuclear energy sources in space (Maharik and Fischhoff, 1993), nuclear weapons (Florig and Fischhoff, 2007), and sexually transmitted infections (Downs et al., 2004).

The core of each application is a normative model informed by descriptive and prescriptive research. How deeply each component is pursued depends on the context of the policy. For some applications, rough normative models suffice; for others, quantitative solutions are needed. Some require dedicated behavioral research; others can rely on existing results, showing general tendencies. Some allow testing prescriptive interventions; others barely invite suggestions for change. Because each example responded to a perceived opportunity (and, sometimes, an actual invitation) to influence policy, they do not represent a well-defined universe of policy choices that depend on assessments of competence. Rather, they illustrate the variety of possible applications and their, sometimes surprising, policy implications.

Applications

Saw Palmetto: Consumers' Competence to Make Decisions Created by Commercial Freedom of Speech

POLICY CONTEXT

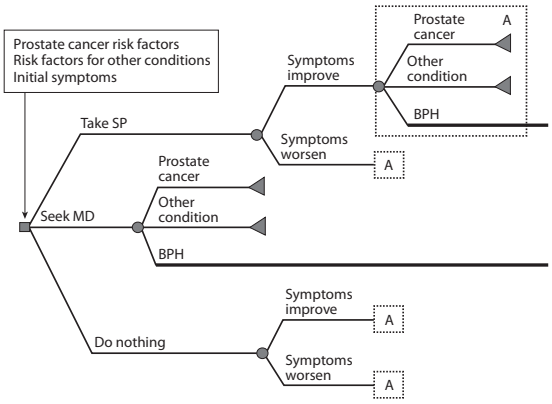
In the United States, the dietary supplements industry has long exceeded the $14 billion estimated a decade ago, with over half of Americans consuming at least one of the 29,000 products (Food and Drug Administration [FDA], 2000a). Currently, the law treats supplements as "reasonably expected to be safe, unless adulterated." The FDA bears the burden of proof for demonstrating harm and for ensuring that label information is "truthful, non-misleading, and sufficient to communicate any risk." Labels can make nonmedical claims without FDA approval but must be withdrawn if the FDA demonstrates that they are *potentially misleading* to a *reasonable consumer* (FDA, 2002; emphasis added). Thus, the policy rests on whether consumer behavior meets a normative standard.

Some supplement manufacturers and consumer advocacy groups have argued that consumers are competent to evaluate such claims. In *Pearson v. Shalala*, the U.S. Court of Appeals accepted this argument, ruling that consumers might benefit from an unverified claim if they knew that they had to evaluate it and could do so. One commentator praised the ruling as ending "FDA's paternalistic approach . . . based on the counterintuitive notion [that] consumers lack the sophistication necessary to evaluate truthful and non-misleading health information" (Emord, 2000, p. 140). If the FDA's "counterintuitive notion" is accurate, though, the ruling denies consumers needed protection.

Eggers and Fischhoff (2004) analyzed this question for a supplement central to the litigation: saw palmetto (SP), a dietary supplement that might relieve lower urinary tract symptoms of benign prostatic hyperplasia (BPH), a chronic problem common among older men.

STEP 1: NORMATIVE ANALYSIS

Figure 12.1 shows the decision facing men with those symptoms. The options (the square *choice* node) are (1) consume the recommended dose of saw palmetto, (2) consult a physician (seek MD), and (3) do nothing. One uncertain *outcome* (triangles) is the change in BPH-related symptoms. Others are the health effects of prostate cancer and other conditions with similar symptoms. The circular *chance* nodes capture the associated uncertainties. The probability of the other conditions should not depend on whether men self-treat with saw palmetto. However, it might take longer to discover those conditions, perhaps long enough to affect the chances of effective treatment (Coley, Barry, and Mulley, 1997).



12.1. Decision tree for men considering saw palmetto (SP) as a treatment for lower urinary track problems. Other decision options (*square node*) are getting medical help (Seek MD) and doing nothing. The primary uncertainty (*circular nodes*) for Take SP and Do Nothing is how symptoms will respond. For Seek MD, it is whether the diagnosis is benign prostatic hyperplasia (BPH), prostate cancer, or some other conditions. The outcomes (*triangles*) depend on the underlying condition and how it responds to treatments, which will depend on how soon it is diagnosed. Box A reflects outcomes with delayed diagnosis.

(Redrawn from Eggers and Fischhoff, 2004)

The scientific evidence on saw palmetto (Marks et al., 2000; Schulz et al., 2002; Wilt et al., 1998) allowed creating a relatively simple decision tree. The tree omits side effects because they are too small to affect choices. It omits costs because they, too, are small ($10/month). It has a single-dose option (the recommended one) because taking more has no known (positive or negative) effects (Ernst, 2002). A "take less" option could be added; it would have a lesser chance of symptom relief and, hence, a greater chance of having other problems diagnosed in time. Eggers and Fischhoff (2004) summarizes that evidence in decision-relevant terms, taking the outcome probabilities from medical evidence and the values from studies of utilities for health states (Tengs and Wallace, 2000). The message that emerged from the normative analysis was "Saw palmetto might be worth a try, but don't neglect other possible sources of your

symptoms." The behavioral analysis asks, in effect, how well men can extract this easily understood message from the information available to them.

STEP 2: DESCRIPTIVE ANALYSIS

The court allowed any product claim, short of preventing or treating specific diseases, if it is accompanied by the disclaimer, "This statement has not been evaluated by the Food and Drug Administration. This product is not intended to diagnose, treat, cure or prevent any disease" (FDA, 2000b). Mason and Scammon (2000) found that people often ignore the disclaimer and, hence, might believe that the FDA had approved the claim. Consumers who notice the disclaimer still might not understand its implications, perhaps assuming that the FDA can regulate dietary supplements as stringently as food (General Accounting Office [GAO], 2000).

Whether such problems are severe enough to undermine people's competency to make these choices is an empirical question. Eggers and Fischhoff (2004) addressed it by asking older men to think aloud as they read, in turn, four labels with increasing detail: (1) *no health claim*, beyond what was inferred from the product name on a green background; (2) an *unqualified claim* of "improving prostate health"; (3) the same health claim, *qualified* by the court-mandated disclaimer; and (4) *full information*, summarizing decision-relevant evidence in a drug fact box (Schwartz, Woloshin, and Welch, 2009; Woloshin, Schwartz, and Welch, 2008). After reading each label, the respondents said whether they would use the product if they had BPH symptoms and for how long if the symptoms persisted. Given the respondents' extensive study of the materials, their choices after reading the fact box were treated as the ones that they should make. For 55%, that choice was to take saw palmetto.

The interview transcripts provide rich qualitative detail regarding the respondents' beliefs about supplements, labeling, and regulation. Most saw the health claim as advertising and the disclaimer as perfunctory (e.g., for liability protection). Many interpreted prostate health as referring to prostate cancer or sexual function. The (unqualified) health claim prompted many to offer higher estimates of positive effects than with full information. Adding the disclaimer decreased the product's perceived efficacy for some respondents while increasing it for others, who gave explanations such as, the "FDA doesn't believe in alternative medicine." The full-information label sometimes increased judged side effects, sometimes reduced them. Almost all respondents said that they would check with their physicians should symptoms persist. Gades et al. (2005) found that about 50% of men with lower urinary tract symptoms seek medical treatment.

STEP 3: PRESCRIPTIVE ANALYSIS

Table 12.1 compares whether respondents *should use* saw palmetto with whether they said that they *would use* it after reading each label. With no claim, all those for whom the optimal choice is taking saw palmetto would miss its potential benefits (row 2). With the unqualified claim, most of them would use it (44% in row 1 vs. 11% in row 2), as would some for whom it was inappropriate (row 3). Adding the disclaimer to the unqualified claim reduces appropriate choices by discouraging some men who should try saw palmetto (row 1).

POLICY ANALYSIS

Table 12.1 predicts the distributions of outcomes with three policies, each embodied by a different label. The policy choice should depend on the weight that regulators assign to each cell. If they weight all cells equally, then they should prefer the unqualified claim, which produces the most appropriate choices (row 1 + row 4). They should prefer no claim, if mistakenly taking the product (row 3) is much worse than missing its potential benefits (row 2). They should never add the disclaimer, which is worse than the (somewhat misleading) unqualified claim.

Consumers are competent to make a decision if the distribution of outcomes under a policy falls within the regulators' tolerances. Here, although the unqualified-claim label dominated the claim + disclaimer label, its outcomes would be acceptable only if the 77% correct choices (row 1 + row 4) outweighed the 22% incorrect ones (row 2 + row 3). Having 11% take saw palmetto inappropriately seems of little

Table 12.1 Predicted optimality of consumer choices with alternative saw palmetto labels

Normative decision	Predicted decision	Label		
Should use	Would use	No claim (%)	Unqualified claim (%)	Claim + disclaimer (%)
Yes	**Yes**	**0**	**44**	**27**
Yes	No	55	11	27
No	Yes	0	11	13
No	**No**	**45**	**33**	**33**

Note: The optimal decision ("should use") reflects the choices of respondents who studied the full-disclosure label. Bold indicates appropriate decisions. See Eggers and Fischhoff, 2004, for details.

consequence, given that almost all respondents said that they would see a doctor if symptoms persisted, thus reducing the opportunity costs of trying saw palmetto. Given its limited efficacy, not much is lost by having another 11% fail to give it a warranted try. By the same logic, even the confusing disclaimer might allow competent choices with this benign product and such moderately engaged, informed, and skeptical consumers.

A product that might produce similar conclusions is black cohosh, a supplement that might reduce menopausal symptoms. Its decision tree resembles that of figure 12.1, with the only major risks arising from delayed treatment. Behavioral research might reveal similar consumer competence, perhaps reflecting heuristics like "Supplements are worth a try, but not for too long, if they're not working" and "Tell your doctors what you're taking, even if you expect them to be skeptical." With other products, though, the disclaimer might exact a high price, if (as seen here) consumers ignore it, brush it off as perfunctory, or have it remind them about the possibilities of alternative medicine. With most dietary supplements, the risks and benefits are unknown. Saw palmetto and black cohosh are unusual in that they have been studied enough to allow estimates of their effects.

Even if a label produces acceptable outcomes, policy makers need not accept it. Schwartz, Woloshin, and Welch (2009) found that most people can understand fact boxes that summarize the risks and benefits of possible treatments, like those used in the full-information label. Consumers are better served by policies that require providing the facts needed to make sound choices, rather than just claims and disclaimers that "work well enough" in forgiving situations. Developing such messages requires analytical and empirical research. It should not be left to intuition, even that of well-intended jurists.

Plan B Morning-After Pill: Adolescents' Competence to Make Reproductive Decisions

POLICY CONTEXT

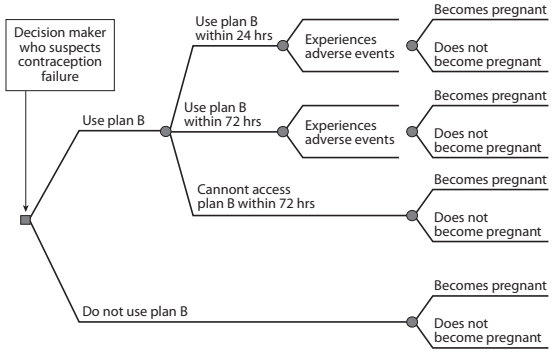
Plan B is an emergency contraceptive (EC) pill that reduces the probability of pregnancy if two doses are taken within 72 to 120 hours of unprotected sex (von Hertzen et al., 2002). In August 2000, the FDA approved the over-the-counter (OTC) sale of EC to women aged 18 and older in outlets that have accredited pharmacies and avoid selling it to younger women. In 2003, the drug's manufacturer, Barr Pharmaceuticals, petitioned the FDA to approve OTC status for all women in less-restricted outlets. Such approval required demonstrating the drug's safety and efficacy without physician supervision (Pub. L. No. 82-215, 65 Stat. 648, 1951). The producer submitted clinical trial data, along with behavioral evidence regarding label comprehension and use, under simulated OTC conditions. An advisory panel (Sherman, 2004) recommended OTC status by a vote of 23 to 4 (FDA, 2003). However, the FDA's Center for Drug Evaluation and Research (CDER), which governs prescription-to-OTC switches, denied approval (FDA, 2004). In his explanatory note, CDER's acting director objected to extrapolating behavioral data from older adolescents to ones under 16, who might exhibit "impulsive behavior, without the cognitive ability to understand the etiology of their behavior" (GAO, 2005, p. 46). He also expressed concern about "the potential impacts that the OTC marketing of Plan B would have on the propensity for younger adolescents to engage in unsafe sexual behaviors due to their possible lack of cognitive maturing compared to older adolescents" (GAO, 2005, p. 5).

The parties to this rule making invoked opposing theories of teen decision-making competence. Critics of OTC status predicted that making Plan B more available would encourage unprotected intercourse, with teens knowing that they can still prevent pregnancy—so that Plan B becomes Plan A. Sexually transmitted infections (STIs) would then increase, even if unplanned pregnancies did not. Some critics worried about increased intercourse among unmarried teens per se, an outcome with no status under the law (FDA, 2003). In contrast, proponents of OTC status argued that teens were sufficiently competent as decision makers that they would not only maintain their current contraceptive practices but also make better choices given the additional option. These advocates also predicted that reducing unintended pregnancies would reduce abortions.

STEP 1: NORMATIVE ANALYSIS

The deliberations over Plan B addressed two decisions within FDA's statutory public health mission: how adolescents choose to protect themselves against (1) unwanted pregnancy and (2) STIs. Some panel members seemingly considered decisions outside the statutes: whether adolescents have sex and abortions. Figure 12.2 analyzes one decision whose options and outcomes clearly fall within the FDA's mandate: whether women use Plan B after suspected contraceptive failure. Krishnamurti, Eggers, and Fischhoff (2008) provide additional details on this choice and two others: whether women have sex and what protection they use, if they do.

Predicting the outcomes in these normative analyses is generally straightforward, because the health



12.2. Decision tree for women considering the use of emergency contraception, contingent on suspecting failure of contraceptive method.
(Redrawn from Krishnamurti, Eggers, and Fischhoff, 2008)

outcomes are often well studied. Evaluating the outcomes is more difficult. Some outcomes are valued similarly by most women and hence might be assessed with standard utility measures; others depend on the woman facing them. In some life situations, women find unplanned pregnancies enormously disruptive; others take them in stride. Some women view pregnancies avoided by Plan B as abortion; others do not. Teen (or unmarried) sex is acceptable to some people, sinful to others. These normative analyses used the values of the women making them. Decisions optimal for them might be unacceptable to people with different values (e.g., regarding the morality of extramarital sex or various birth-control options).

STEP 2: DESCRIPTIVE ANALYSIS

Krishnamurti, Eggers, and Fischhoff conducted (2008) semistructured, open-ended interviews with 30 adolescents from high-risk populations whose ages straddled 16 years old (critical to the FDA's decision). The interviews followed a mental-models protocol (Morgan et al., 2001), which directs respondents increasingly toward issues in the normative analysis. Such interviews have the statistical power to reveal large group differences. Their results guide the design of structured surveys suited to larger samples (which confirmed the results reported here).

As with any other study, the results of these interviews should be interpreted in the context of general scientific knowledge. In this case, a broad summary (Fischhoff, 2008; Reyna and Farley, 2006) might be that, by the midteen years, most adolescents have roughly the cognitive decision-making skills of adults. The GAO (2005) noted that the FDA's Plan B decision departed from earlier rulings, consistent with research, that had found it "scientifically appropriate to extrapolate data from older to younger adolescents" (p. 5).

Whatever their skills, teens' knowledge of a topic depends on their opportunities to learn about it through exposure to information (e.g., instruction, media, word-of-mouth) and experience. Krishnamurti, Eggers, and Fischhoff found that their respondents generally said that they (1) knew about Plan B; (2) would not use it as their primary birth-control method (usually citing discomfort or cost); (3) would consider it if they suspected an unplanned pregnancy; (4) saw barriers to access that expanded OTC status would reduce; and (5) had considered whether Plan B constituted abortion. Based on these accounts, Krishnamurti, Eggers, and Fischhoff concluded that expanded OTC status would not affect the sexual behavior or the choice of primary birth-control method; however, it would increase Plan B use when needed. They also found considerable confusion about the timing of usage, with more teens underestimating the effective period. The only observed difference between teens over and under 16 was that younger teens saw greater costs and fewer benefits with Plan B and hence were less likely to use it.

STEP 3: PRESCRIPTIVE ANALYSIS

Based on these interviews, expanded OTC availability would affect only one decision: Plan B use after unprotected sex. Therefore, it would allow some young women to achieve a desired end that is currently unfeasible. Although the teens' decision-making processes were thoughtful, they were not always well informed, particularly regarding Plan B's effective period. Women who underestimate its effectiveness will not receive the full benefit from greater availability. Those who overestimate it will not receive the benefits that they expect.

POLICY ANALYSIS

In terms of the health outcomes in the FDA's regulatory mandate, these results suggest that expanded OTC availability dominated limited availability: There should be fewer unplanned pregnancies, with no changes in other health effects. The largest gap in teens' knowledge, the period of EC effectiveness, should not be hard to close with a suitable communication channel. Emotion might, in principle, undermine the cognitive competencies seen here (and in other studies). In practice, though, it seems that OTC availability should reduce emotional pressure in the one decision where EC seems to play a role: dealing with unplanned, unprotected sex.

Going beyond the FDA's mandate allows the consideration of abortion-related issues. In the study,

most respondents who were opposed to abortion thought that Plan B was used too soon after intercourse to constitute abortion. For them, expanded availability would allow more optimal choices. That conclusion should be rejected by individuals who oppose abortion and place conception at intercourse. They should fault such teens' beliefs (about how Plan B works), but not their cognitive abilities.

Carotid Endarterectomy: Medical Informed Consent

POLICY CONTEXT

In the United States, roughly half of the states hold physicians to a *materiality* standard when securing informed consent. That is, they must ensure that patients understand all of the facts material to the decision. (The other states have a *professional* standard, requiring adherence to common practice.) That standard implies a normative analysis—ordering information by its materiality—that dictates when physicians can stop communicating because they have conveyed all that patients need to know. Having that standard should provide some protection against unfair legal judgments when things go wrong and patients claim that they were not adequately informed. Merz et al. (1993) offered an approach to formalizing the materiality standard, which they illustrate with a common surgical procedure whose large risks and benefits have engendered extensive research on expected outcomes.

Carotid endarterechtomy, scraping out the artery to the brain, can reduce the risk of stroke for patients with atherosclerosis. However, many things can go wrong, ranging from death and iatrogenic strokes to headaches and broken teeth. Some risks are specific to the procedure; others are those of major surgery. The full list is a lot to consider when facing difficult tradeoffs between these risks and the expected benefits of treating a life-threatening illness. Informed consent is a vague policy that does not specify what patients need to know.

STEP 1: NORMATIVE ANALYSIS

This method creates hypothetical patients for whom surgery would be the optimal choice were there no risks (and were money no object). These patients vary in their physical condition, as represented by probability distributions over possible outcomes, and in their personal values, as represented by utility distributions over the outcomes. Patients are created by sampling values from these distributions (with the simplifying assumption of independence in this application). The expected utility of surgery is calculated for each such patient, ignoring all risks. It is positive, given how the population is created. The expected utility is then recalculated, incorporating knowledge of each possible side effect. A side effect's materiality is defined as how often it gives the surgery negative expected utility, so that it is no longer recommended.

Merz et al. found that only three side effects should matter to many of these simulated patients: about 15% should decline surgery if told of the risk of immediate death; another 5% if told of the risk of iatrogenic stroke; and 3% more if told of the risk of facial paralysis. Learning about other risks would tip the scales for very few patients.

STEP 2: DESCRIPTIVE ANALYSIS

Because few candidates for this surgery have faced it before, they cannot be expected to have authoritative risk information. As a result, Merz et al. (1993) conducted no behavioral research, assuming that all information would have to be conveyed. This strategy would be flawed if patients held strong beliefs about other risks that communications about the three focal risks would ignore (Downs, Bruine de Bruin, and Fischhoff, 2008). In cases where there are many material facts, assessing current beliefs allows identifying those that go without saying—and can be skipped in favor of new ones.

STEP 3: PRESCRIPTIVE ANALYSIS

Thus, effectively conveying the three major side effects would keep 15%–20% of these potential patients from undergoing surgery that is suboptimal for them (because its risks outweigh its benefits). Based on risk-communication research (Fischhoff, 2009; Schwartz et al., 2009), that task seems manageable. Each critical side effect involves a fairly familiar event (death, stroke, facial paralysis) with a fairly large probability (not, say, 1 in 46,000). Thus, it should be possible to afford patients the knowledge needed for competent decision making. A greater prescriptive challenge may be helping patients to be rational enough to take advantage of that information. The empirical-analytical question is how sensitive these choices are to imperfect information integration (von Winterfeldt and Edwards, 1986).

POLICY ANALYSIS

If the materiality standard is implemented in this way, the duty to inform is fulfilled once the three critical side effects are communicated. Doing so absolves physicians of responsibility for suboptimal choices that reflect poor information integration rather than poor understanding. Focusing communication on the most material facts should improve patients' information integration by protecting them from full-disclosure

practices that drown them in immaterial facts. Full disclosure is needed to perform the analyses that identify the critical risks (and to ensure that nothing is hidden); however, forcing patients to sort through all the details can undermine their competence.

Materiality analysis can also help to implement research-funding policies by assessing the value of potential results for patient decision making. Those analyses could help patients decide whether to wait for clinical trial outcomes.

Methylene Chloride–Based Paint Stripper: Consumer Competence to Make Decisions Created by Voluntary Self-Regulation

POLICY CONTEXT

In the mid-1990s, the International Agency for Research on Cancer declared the solvent methylene chloride a probable carcinogen. Methylene chloride also gives off carbon monoxide, which can cause heart attacks in confined spaces. One use of the solvent with no clear substitute is that of stripping paint. Although industrial users might be required to take protective measures (e.g., ventilating hoods, respirators), home users must protect themselves. If they can, then it is safe to leave the paint stripper on the market. That depends on their competence to make decisions about how to use the product and about whether it has acceptable risks relative to its benefits.

STEP 1: NORMATIVE ANALYSIS

Riley et al. (2001) characterized consumers' health risks with an analysis sensitive to their usage patterns. It predicted consumer exposures based on physical principles (e.g., air circulation, chemistry), whose interactions had been calibrated under laboratory conditions, and on the effects of actions that users might decide to take (e.g., open windows, wait in corner while stripper is curing). The analysis created a *supply curve*, sorting the actions in decreasing order of marginal effectiveness for reducing exposure. That order is the logical one for adopting risk-control measures (assuming similar costs). The analysis found that for many jobs (varying in room size, duration, etc.), consumers could greatly reduce their risks by deciding to implement two simple actions: opening a window and having a fan blow outward.

STEP 2: DESCRIPTIVE ANALYSIS

Interviews with paint-stripper users found that they (1) were motivated to use control measures, (2) could easily understand the two key measures, and (3) seemed realistic about their ability to execute them (e.g., "I'm not going to open a window in the winter, however bad the fumes"). However, those actions were sufficiently unintuitive that consumers needed to have them explained. For example, without instruction, many would have had the fans blow inward, in order to feel the airflow, not realizing that internal circulation left concentrations of the chemical unchanged. Many also reported choosing actions whose ineffectiveness could be easily explained (e.g., gloves get ruined because the solvent dissolves them; fumes diffuse, so do not bother crossing the room while the solvent cures). Thus, users appeared competent to make sound choices when provided with the relevant information.

STEP 3: PRESCRIPTIVE ANALYSIS

In lieu of in-home observation of consumers' usage decisions, Riley et al. (2001) estimated the exposures for users who understood and conscientiously followed everything they read but who had read different labels and reading patterns (e.g., just the instructions, just the warnings, just the bolded material, just the first five items). The analyses found wide variation in exposures across labels and reading patterns. Some labels provided useful information whatever the reading strategy; some had contained no information on reducing exposures. Thus, consumers are competent to make sound usage decisions if they receive relevant information about how to use the product. Without that information, they cannot know what risks they are taking or whether to use the product. Predictions of actual label-reading behavior and the resultant risk levels could use direct observation or general patterns (Wogalter, 2006).

POLICY ANALYSIS

Flawed communications limit otherwise competent consumers' ability to make effective choices. As a result, they may purchase unduly risky products and use them in needlessly risky ways, or, consumers may forego useful products and take needless precautions. When a product's risks and benefits depend on how it is used, its label is as much a part of the product as are its physical constituents. Regulators should want to know what consumers take away from labels, so that they can ensure proper protections. Producers should want the same knowledge, so that they can help consumers get the greatest value from their products and defend themselves against charges of failing to fulfill their duty to inform. Individual firms produced the labels in this study. The great variation in the value of the information that they provided (and the associated

risks) suggests that such voluntary self-regulation was inadequate.

Cryptosporidium: Consumer Competence to Cope with a Contamination Emergency

POLICY CONTEXT

Cryptosporidium is a common protozoan parasite that has mammalian hosts and can infect public water supplies, typically through uncontrolled sewage discharges and fecally contaminated runoff after heavy storms (e.g., from feedlots, deer). Typical water treatment systems cannot fully remove or deactivate it. Symptoms, which appear 1 to 7 days after exposure, include nausea, vomiting, diarrhea, cramps, and low fever. Although cryptosporidiosis has no medical cure, most infected individuals recover, many without exhibiting symptoms. However, the disease, which attacks the liver, can be fatal to immunocompromised individuals (e.g., those with AIDS). Water and public health authorities have a duty to inform consumers so that they can make competent water-usage decisions, namely when and how to use boiled or bottled water.

STEP 1: NORMATIVE ANALYSIS

Casman et al. (2000) created a model predicting the health effects of a *Cryptosporidium* intrusion. The model includes inputs from microbiology (dose-response relationships), civil engineering (filtration, testing), ecology (upstream land use), communication research (dissemination of "boil water" notices), and psychology (perceived risk, actual response). It allows assessing of when consumers are competent to make water-usage decisions under the conditions created by different intrusion scenarios. Among other things, it examines how quickly they receive messages, how adequately they boil water, and how much they rely on personal testing (which has little value).

STEP 2: DESCRIPTIVE ANALYSIS

The model uses estimates from observational studies of consumers' water-use decisions in past intrusions, which have found that people often use improperly treated water. New interviews found that the main sources of these poor decisions were ignorance (about how to boil water well enough to destroy the parasite) and suspicion (about how seriously to take warnings). Both the procedures and the context are simple enough that it should be possible to explain them well enough to allow most people to make competent choices. As a result, the competence of consumers to choose depends on whether they get sound messages in time to act.

STEP 3: PRESCRIPTIVE ANALYSIS

The model was used to predict health effects under the conditions created by plausible intrusion scenarios that differed, for example, in how long it took to detect and repair problems. One set of simulations examined the effects of ensuring that all consumers received sound messages as soon as an intrusion was established. Such communication was found to have no effect, a result that was traced to *Cryptosporidium* testing procedures being too slow for messages to arrive in time to prevent exposures. As a result, relying on "boil water" notices ensures suboptimal choices by individuals who could be competent given the right information.

POLICY ANALYSIS

In this case, the analysis revealed an inherent flaw in a standard policy, relying on consumer decision making to manage risks—a faith that deflects attention from other possible policies. One other possibility is improving the speed of detection, so that officials can provide timely warning. Another possibility is reducing the risk, through better land use or water purification. A third possibility is routinely providing highly vulnerable populations with safe water. Repeated with contaminants allowing rapid detection (e.g., some *E. coli* strains), the analysis might reveal that timely, comprehensible warnings would allow effective consumer decision making. Consumers should not be blamed for negative outcomes when they cannot help themselves.

Emergency Evacuation: Citizen Competence to Make Voluntary Decisions

POLICY CONTEXT

Mass emergencies can be defining moments for a society (Boin et al., 2005), ones in which leaders' decision-making competence will be examined intensely. One aspect of that examination will be how their actions affected their citizens' decision making. Were citizens afforded the information needed to make effective choices? Did they have the resources needed to act on that information? Were they treated like competent adults? Was martial law imposed when they could have managed without it?

Emergency plans must make some assumptions about the public's decision-making competence under stressful conditions. One such class of hazards

is terror attacks with contaminating materials, such as radioactive dispersion devices (RDDs), or dirty bombs. RDDs use ordinary explosives to spread radioactive materials, causing immediate casualties from the blast, long-term casualties from radiation poisoning, and potentially great social and economic costs. The extent of that disruption will depend on their leaders' perceived competence, which will, in turn, depend on how well those leaders assess their public's competence. If they expect too much, then the public will be denied needed protection. If they expect too little, the public will be denied deserved freedom.

STEP 1: NORMATIVE ANALYSIS

Dombroski and Fischbeck (2006) and Dombroski, Fischhoff, and Fischbeck (2006) developed a general model for predicting the health effects of RDD attacks, with one predictor being citizens' decisions about evacuating or sheltering in place. The model incorporates features from research into explosive impacts, aerosol dispersion, traffic flows, dose-response relationships, and so on. Choosing values for the model parameters (e.g., location, time of day, explosive force, contaminant, weather) produces attack scenarios specific enough to predict morbidity and mortality (from which economic and social effects might be predicted). For public health officials, the optimal response minimizes those health effects. One key decision is whether to recommend evacuation or sheltering in place. For individual citizens, the optimal response minimizes those risks, subject to other concerns (e.g., protecting family members, helping coworkers, demonstrating resilience). For them, one key decision is whether to follow that recommendation.

STEP 2: DESCRIPTIVE ANALYSIS

Estimates for most model parameters were taken from the research literature. However, although there are many studies of emergency behavior, they are rarely in model-ready form. Therefore, judgments of the model's behavioral parameters were elicited from 10 social science experts and 36 local disaster specialists. These experts predicted the behavior for variations of a scenario involving a 10-kilogram Cs-137 RDD exploded at Pittsburgh's USX Tower at 10 a.m. on a summer weekday. Their judgments included the percentages of citizens complying with instructions to evacuate or shelter in place when at home and at work. These experts generally agreed about the citizens' decisions. Consistent with historical experience, although not with popular myth (Tierney, Lindell, and Perry, 2001; Wessely, 2005), the experts expected no panic. Instead, they expected most people to follow the instructions, with higher rates for sheltering at home and evacuating from work.

STEP 3: PRESCRIPTIVE ANALYSIS

Incorporating the experts' judgments in the model revealed that for this scenario, the predicted rates of compliance with official instructions (60%–80%) are good enough to minimize the health effects. Enough people would shelter in place to keep the roads open enough to allow first responders to treat those injured in the blast while not trapping evacuees in a radioactive cloud. Little would be gained by making the recommendation compulsory.

POLICY ANALYSIS

These results suggest that citizens are competent enough to make choices that achieve generally optimal outcomes. Thus, voluntary compliance should satisfy a consequentialist regulatory philosophy. It should have additional procedural value for demonstrating faith in the public compared to compulsory policies, like martial law. That value underlies the commitment to "keep the public fully informed—tell what we know, tell what we don't know, and tell it often. . . . Maintain credibility and public trust, by providing accurate, science-based information" (Department of Health and Human Services, 2006). Producing and disseminating useful information is one way to earn trust.

Strategies for Competence Assessment

Reprise

More or less the same kinds of (ordinary) people with more or less the same general decision-making skills were involved in each of these examples. Yet, their competence emerges differently in each, depending on the difficulty of the choice and the adequacy of others' attempts to fulfill a duty to inform. Normative, descriptive, and prescriptive analyses allowed evaluation of the proposed policies in terms of how well they fit the competence that individuals bring to them and to design better ones.

According to these analyses:

With saw palmetto, the policy of allowing any nonhealth claim if accompanied by a court-mandated disclaimer created decisions that consumers were competent to make despite the disclaimer's flaws, as a result of the product being benign and the consumers being skeptical. Taking saw palmetto should not hurt them or lead them to delay medical care too

long. The disclaimer's main impact was leading some consumers to forego a product that might help them. With other products and consumers, the policy might have much worse effects.

With Plan B, young women should be competent to make the decisions created by the policy of expanded OTC availability. It should help them to avoid unwanted pregnancies without violating their abortion views. Expanded availability should not affect their decisions about sexual behavior or contraceptives.

With carotid endarterechtomy, most patients should be competent to make the decisions created by a policy of focusing patient briefings on the most material facts. In states with a materiality standard, conveying those few, simple facts might allow physicians to claim to have secured informed consent.

With methylene chloride paint stripper, most consumers would not be competent to make the decisions created by the current policy of allowing producers to design their own labels. However, the critical facts are simple: decide to use the product only if you can have a fan blow outward through an open window. Thus, a policy that mandated labels with that information should allow competent choices.

With *Cryptosporidium*, consumers are not competent to make water usage decisions under the conditions created by a policy of relying on them to protect themselves because critical information will not arrive in time. In effect, such a policy asks consumers to do the impossible whatever their decision-making skills. It might allow them to make competent choices about contaminants if faster testing were available.

With the RDD scenario, citizens are competent to decide whether to obey recommendations to evacuate or shelter in place. Trusting that competence should enhance citizens' trust in their authorities. Thus, a policy of making clear recommendations should dominate draconian policies like martial law.

Organizing for Assessing (and Improving) Decision-Making Competence

Sweeping claims about decision-making competence cannot do justice to the diversity of decisions and decision makers. Blanket claims of competence leave some people without needed protections, and blanket claims of incompetence deprive some of deserved freedoms. Blanket claims create the temptation of working backward from desired policies to the behavioral assumptions that justify them. An incompetent public suits those who favor strong regulations and technocratic management. A competent public suits those who favor free markets and participatory processes. A disciplined approach, combining empirical and analytical research, is needed to determine the legitimacy of such policies.

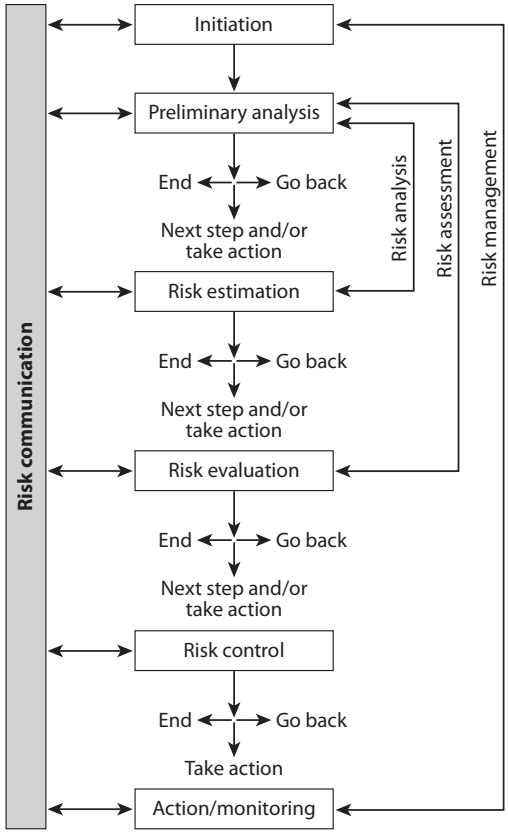
Executing this approach to assessing (and improving) decision-making competence requires an interdisciplinary team, with

1. *Subject-matter experts*, who are able to identify the decision options and characterize the processes determining their effects;
2. *Decision analysts*, who are able to estimate the risks and benefits of those options, showing the facts most relevant to decision making—and competence;
3. *Social scientists*, who are able to assess decision makers' beliefs and values, guide attempts to enhance competence, and evaluate their success; and
4. *Designers*, who are able to implement measures that achieve theoretically possible competence.

Assembling such a team requires leadership. Often, policy-making organizations are dominated by experts from one field who have little interest in collaborating with others. They may want to avoid sharing scarce resources. They may not recognize the limits to their own expertise. When such organizations expand, they face the challenge of evaluating unfamiliar expertise. If they cannot tell what "good" is and select poor representatives of another field, they may both get poor advice and devalue that field.

Once present, these experts must be coordinated. A strong team will accept ideas from anyone; however, it must assign responsibility for each task to the appropriate experts. Subject-matter experts can best predict the outcomes that concern decision makers, but social scientists are needed to assess what the effects of those outcomes will be. Subject-matter experts will know which facts are central to their professional community, but decision analysts are needed to determine their relevance. System designers will know how to get messages out, but social scientists are needed to determine how well their content is absorbed. Social scientists will know how difficult the tasks are, but decision analysts are needed to assess how sensitive the decisions are to those effects.

Once assembled and coordinated, the competence-assessment team must be integrated into the policy-making process. Figure 12.3 shows an organizational model with that goal. Taken from the quasi-governmental Canadian Standards Association (1997), it is consistent with recommendations from Her Majesty's Treasury (2005), the National Research Council (1996), and the Presidential/Congressional Commission on Risk Assessment and Risk Management (1997), among other bodies. The center of the figure depicts a standard policy-making process, unusual only in that it evaluates each stage



12.3. Steps in the Q850 Risk Management Decision-Making Process—simple model. *Note*: Risk communication with stakeholders is an important part of each step in the decision process.

(Redrawn from Canadian Standards Association, 1997)

before proceeding to the next (with the implicit possibility of never finishing). Notably, each stage requires two-way risk communication. For example, the initiation stage entails experts learning which issues matter to those whom a policy affects and telling them how those issues will be addressed. These communications might be direct or indirect, with social researchers soliciting views and conveying results.

Such a process allows considering decision-making competence early enough to shape the design of policies and allow midcourse corrections. It invites recruiting the kinds of expertise needed to make the work *behaviorally realistic* in its assumptions about individuals' ability to secure, comprehend, and use information and *analytically sound* in its sensitivity to the heterogeneity in people's abilities and decisions. Without such integrated expertise, it is impossible to do justice to individuals' needs and limitations and to create policies that afford them as much autonomy as they want and can handle.

Note

Preparation of this chapter was supported by U.S. National Science Foundation Grant SES 0433152 and the Veterans Administration Center for Health Equity Research and Policy. We thank Wändi Bruine de Bruin, Julie Downs, Irene Janis, Valerie Reyna, and two anonymous reviewers. The views expressed are those of the authors.

References

Boin, A., 't Hart, P., Stern, E., and Sundelius, B. (2005). *The politics of crisis management: Public leadership under pressure*. Cambridge: Cambridge University Press.

Bruine de Bruin, W., Parker, A., and Fischhoff, B. (2007) Individual differences in adult decision-making competence (A-DMC). *Journal of Personality and Social Psychology*, *92*, 938–956.

Camerer, C., Issacharoff, S., Loewenstein, G., O'Donoghue, T., and Rabin, M. (2003). Regulation for conservatives: Behavioral economics and the case for "asymmetric paternalism." *University of Pennsylvania Law Review*, *151*, 1211–1254.

Canadian Standards Association. (1997; reaffirmed 2009). *Risk management: Guidelines for decision makers*. (CSA-850). Ottawa: Canadian Standards Association.

Casman, E., Fischhoff, B., Palmgren, C., Small, M., and Wu, F. (2000). An integrated risk model of a drinking-water–borne Cryptosporidiosis outbreak. *Risk Analysis*, *20*, 493–509.

Coley, C. M., Barry, M. J., and Mulley, A. G. (1997). Screening for prostate cancer. *Annals of Internal Medicine*, *126*, 480–484.

Department of Health and Human Services. (2006). DHHS communication plan for first case of H5N1 in US. Washington, DC: Department of Health and Human Services.

Dombroski, M., Fischhoff, B., and Fischbeck, P. (2006). Predicting emergency evacuation and sheltering behavior: A structured analytical approach. *Risk Analysis*, *26*, 501–514.

Dombroski, M. J., and Fischbeck, P. S. (2006). An integrated physical dispersion and behavioral response model for risk assessment of radiological dispersion device (RDD) events. *Risk Analysis, 26*, 501–514.

Downs, J. S., Bruine de Bruin, W., and Fischhoff, B. (2008). Patients' vaccination comprehension and decisions. *Vaccine*, *26*, 1595–1607.

Downs, J. S., Murray, P. J., Bruine de Bruin, W., Penrose, J., Palmgren, C. and Fischhoff, B. (2004). Interactive

video behavioral intervention to reduce adolescent females' STD risk: A randomized controlled trial. *Social Science and Medicine*, *59*, 1561–1572.

Edwards, W. (1954). The theory of decision making. *Psychological Bulletin*, *51*(4), 380–417. doi:10.1037/h0053870

Eggers, S. L., and Fischhoff, B. (2004). Setting policies for consumer communications: A behavioral decision research approach. *Journal of Public Policy and Marketing*, *23*, 14–27.

Emord, J. W. (2000). *Pearson v. Shalala*: The beginning of the end for FDA speech suppression. *Journal of Public Policy and Marketing*, *19*(1), 139–143.

Ernst, E. (2002). The risk-benefit profile of commonly used herbal therapies: Ginkgo, St. John's Wort, ginseng, echinacea, saw palmetto, and kava. *Annals of Internal Medicine*, *136*(1), 42–53.

Fischhoff, B. (1992). Giving advice: Decision theory perspectives on sexual assault. *American Psychologist*, *47*, 577–588.

———. (2008). Assessing adolescent decision-making competence. *Developmental Review*, *28*, 12–28.

———. (2009). Risk perception and communication. In R. Detels, R. Beaglehole, M. A. Lansang, and M. Gulliford (Eds.), *Oxford textbook of public health* (5th ed., pp. 940–952). Oxford: Oxford University Press.

Fischhoff, B., Atran, S., and Fischhoff, N. (2007). Counting casualties: A framework for respectful, useful records. *Journal of Risk and Uncertainty*, *34*, 1–19.

Fischhoff, B., Bruine de Bruin, W., Guvenc, U., Caruso, D., and Brilliant, L. (2006). Analyzing disaster risks and plans: An avian flu example. *Journal of Risk and Uncertainty*, *33*, 133–151.

Florig, K., and Fischhoff, B. (2007). Individuals' decisions affecting radiation exposure after a nuclear event. *Health Physics*, *92*, 475–483.

Food and Drug Administration. (2000a). FDA to Jonathan W. Emord, May 26. Retrieved from http://www.fda.gov/ohrms/dockets/dockets/04p0059/04p-0059-pdn0001-03-vol14.pdf

———. (2000b). Regulations on statements made for dietary supplements concerning the effect of the product on the structure or function of the body. 65(4) Fed. Reg. 999 (January 6, 2000) (to be codified at 21 C.F.R. pt. 101)

———. (2002). *Guidance for industry: Structure/function claims small entity compliance guide*. Retrieved from http://www.fda.gov/Food/GuidanceComplianceRegulatoryInformation/GuidanceDocuments/DietarySupplements/ucm103340

———. (2003). *Nonprescription Drugs Advisory Committee in joint session with the Advisory Committee for Reproductive Health Drugs* (meeting transcript). Retrieved from http://www.fda.gov/ohrms/dockets/ac/03/transcripts/4015T1.htm

———. (2004). *FDA's decision regarding Plan B: Questions and answers.* Retrieved from http://www.fda.gov/cder/drug/infopage/planB/planBQandA.htm

Gades, N. M., Jacobson, D. J., Girman C. J., Roberts, R. O., Lieber, M. M., and Jacobsen, S. (2005). Prevalence of conditions potentially associated with lower urinary tract symptoms in men. *BJU International*, *95*, 549–553.

General Accounting Office. (2000). *Food safety: Improvements needed in overseeing the safety of dietary supplements and "functional" foods.* GAO/RCED-00-156. Retrieved from http://www.gao.gov/new.items/rc00156.pdf

———. (2005). *Decision process to deny initial application for OTC Marketing of the emergency contraceptive drug Plan B was unusual.* GAO-06-109. Retrieved from http://www.gao.gov/new.items/d06109.pdf

Gilovich, T., Griffin, D., and Kahneman, D. (Eds.). (2002). *The psychology of judgment: Heuristics and biases.* New York: Cambridge University Press.

Her Majesty's Treasury. (2005). *Managing risks to the public.* London: HM Treasury.

Krishnamurti, T. P., Eggers, S. L., and Fischhoff, B. (2008). The effects of OTC availability of Plan B on teens' contraceptive decision-making. *Social Science and Medicine*, *67*, 618–627.

Maharik, M., and Fischhoff, B. (1993). Public views of using nuclear energy sources in space missions. *Space Policy*, *9*, 99–108.

Marks, L. S., Partin, A. W., Epstein, J. I., Tyler, V. E., Simon, I., et al. (2000). Effects of a saw palmetto herbal blend in men with symptomatic benign prostatic hyperplasia. *Journal of Urology*, *163*, 1451–1456.

Mason M. J., and Scammon, D. L. (2000). Health claims and disclaimers: Extended boundaries and research opportunities in consumer interpretation. *Journal of Public Policy and Marketing*, *19*, 144–50

Merz, J., Fischhoff, B., Mazur, D. J., and Fischbeck. P. S. (1993). Decision-analytic approach to developing standards of disclosure for medical informed consent. *Journal of Toxics and Liability*, *15*, 191–215.

Morgan, M., Fischhoff, B., Bostrom, A., and Atman, C. (2001). *Risk communication: The mental models approach.* New York: Cambridge University Press

National Research Council. (1996). *Understanding risk.* Washington, DC: National Academy Press.

Parker, A., and Fischhoff, B. (2005). Decision-making competence: External validity through an individual-differences approach. *Journal of Behavioral Decision Making*, *18*, 1–27.

Pearson v. Shalala. 164 F.3d 650 (D.C. Cir.) (1999).

Presidential/Congressional Commission on Risk Assessment and Risk Management. (1997). *Final report of the Presidential/Congressional Commission on Risk Assessment and Risk Management. Vol. 2. Risk assessment and risk management in regulatory decision-making.*

Retrieved from http://www.riskworld.com/Nreports/1997/risk-rpt/volume2/pdf/v2epa.PDF

Reyna, V. F., and Farley, F. (2006). Risk and rationality in adolescent decision making. *Psychological Science in the Public Interest*, *7*(1), 1–44.

Riley, D. M., Fischhoff, B., Small, M., and Fischbeck, P. (2001). Evaluating the effectiveness of risk-reduction strategies for consumer chemical products. *Risk Analysis*, *21*, 357–369.

Roper v. Simmons. 543 U.S. 551, 607–608. (2005) (Salia, J. dissenting)

Schulz, M. W., Chen, J., Woo, H. H., Keech, M., Watson, M. E., and Davey, P. J. (2002). A comparison of techniques in patients with benign prostatic hyperplasia. *Journal of Urology*, *168*, 155–159.

Schwartz, L. M., Woloshin, S., and Welch, H.C.G. (2009). Using a drug facts box to communicate drug benefits and harms. *Annals of Internal Medicine*, *150*, 516–527.

Sherman, L. A. (2004). Looking through a window of the FDA. *Preclinica*, *2*(2), 99–102.

Tengs, T. O., and Wallace, A. (2000). One thousand health-related quality-of-life estimates. *Medical Care*, *38*, 583–637.

Tierney, K. J., Lindell, M., and Perry, R.W. (2001). *Facing the unexpected: Disaster preparedness and response in the US*. Washington, DC: National Academy Press.

von Hertzen, H., Piaggio, G., Ding, J., Chen, J., Song, S., Bartfai, G., et al. (2002). Low dose mifepristone and two regimens of levonorgestrel for emergency contraception: A WHO multicentre randomised trial. *Lancet*, *360*, 1803–1810.

von Winterfeldt, D., and Edwards, W. (1986). *Decision analysis and behavioral research*. Cambridge: Cambridge University Press.

Wessely, S. (2005). Don't panic! Short- and long-term psychological reactions to the new terrorism. *Journal of Mental Health*, *14*(1), 106.

Wilt, T., Ishani, A., Stark, G., MacDonald, R., Lau, J., and Mulrow, C. (1998). Saw palmetto extracts for treatment of benign prostatic hyperplasia. *Journal of the American Medical Association*, *280*, 1604–1609.

Wogalter, M. (Ed.). (2006). *The handbook of warnings*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Woloshin, S., Schwartz, L., and Welch, G. H. (2009). *Know your chances*. Berkeley, CA: University of California Press.

CHAPTER 13

If Misfearing Is the Problem, Is Cost-Benefit Analysis the Solution?

CASS R. SUNSTEIN

Many people have argued for cost-benefit analysis on economic grounds. In their view, a primary goal of regulation is to promote economic efficiency, and cost-benefit analysis is admirably well suited to that goal. Arguments of this kind have been met with sharp criticism from those who reject the efficiency criterion or who believe that in practice, cost-benefit analysis is likely to produce a kind of regulatory paralysis ("paralysis by analysis") or to represent a bow in the direction of well-organized private groups.

In this chapter, I offer support for cost-benefit analysis not from the standpoint of conventional economics, but on grounds associated with cognitive psychology and behavioral economics. My basic suggestion is that cost-benefit analysis is best defended as a means of responding to the general problem of *misfearing*, which arises when people are afraid of trivial risks and neglectful of serious ones. For the purposes of law and policy, the central points are twofold. First, predictable problems in individual and social cognition lead citizens and representatives to misfear. Second, misfearing plays a substantial role in public policy, in part because of the power of self-interested private groups and in part because of the ordinary political dynamics. When misallocations of public resources result from misfearing and associated problems, cost-benefit analysis can operate as a corrective. Thus understood, cost-benefit analysis is a way of ensuring better priority setting and of overcoming predictable obstacles to desirable regulation.

Of course much of the controversy over cost-benefit analysis stems from the difficulty of specifying, with particularity, what that form of analysis entails. An understanding of misfearing cannot support any particular understanding of cost-benefit analysis. Certainly I do not mean to embrace the controversial, and indeed implausible, proposition that all regulatory decisions should be made by aggregating private willingness to pay, as if economic efficiency is or should be the goal of all regulation. I will attempt instead to provide a defense of cost-benefit analysis, rooted in cognitive considerations, that is agnostic on large issues of the right and the good and that should be able to attract support from people with diverse theoretical commitments or with uncertainty about the appropriate theoretical commitments.

Misfearing and the Public Demand for Regulation

When people fall prey to misfearing, why, exactly, do they do so? I shall offer several answers, but it will be helpful to orient those answers under a simple framework. A great deal of recent work has stressed two families of cognitive operations in the human mind, sometimes described as System I and System II, through which risky activities and processes are evaluated. System I is fast, associative, and intuitive; System II is more deliberative, calculative, slower, and analytic. The two systems may well have different locations in the brain. But the distinction between the two systems is useful whether or not identifiable brain sectors are involved. The central point is that people have immediate and often visceral reactions to persons, activities, and processes, and the immediate reaction operates as a mental shortcut for a more deliberative or analytic assessment of the underlying issues. Sometimes the shortcut can be overridden, or corrected, by System II. For example, System I might lead people to be terrified of flying in airplanes or of large dogs, but System II might create a deliberative check, ensuring an eventual conclusion that the risks are trivial.

My suggestion here is that misfearing is often a product of System I and that cost-benefit analysis can operate as a kind of System II corrective, ensuring that people have a better sense of what is actually at stake. Of course System II can itself go badly wrong: the analysis of effects may be erroneous, and

the translation of risks into monetary equivalents creates many problems. It would be foolish to contend that System II is error free. The only claims are that System I is prone to make systematic errors, that those errors produce misfearing, and that an effort to assess the costs and benefits of risk reduction, if done properly, will operate as a helpful constraint.

These points work most naturally for individual judgments. If people are afraid of flying and if they are not afraid of smoking cigarettes, a kind of informal cost-benefit analysis might help. The political process is hardly a simple reflection of System I, even in the most responsive democracies. When a legislature or administrative agency is moved to act, a complex dynamic is responsible, and the dynamic prominently includes the activities of self-interested private groups with strong incentives to obtain relevant information. When the public demand for law produces excessive reactions to small risks (Sunstein, 2002), legal rules must surmount a series of barriers to ill-considered public action. Nonetheless, it is clear that public misfearing, often bred by self-interested or altruistic private groups, helps to produce significant misallocations of public resources (for examples, see Sunstein, 2002). To the extent that misfearing is a result of social interactions and political influences, and to the extent that misallocations are a product of a multitude of factors, the argument for cost-benefit analysis is strengthened rather than weakened.

These claims raise an immediate question: What, exactly, is cost-benefit analysis? For the moment, let us understand that approach to require regulators to identify, and to make relevant for purposes of decision, the good effects and the bad effects of regulation and to quantify these as much as possible in terms of both monetary equivalents and life-years saved, hospital admissions prevented, workdays gained, and so forth. Let us also assume that cost-benefit analysis can accommodate distributional factors, by, for example, giving special weights to adverse effects on disadvantaged social groups. How might cost-benefit analysis help to correct the problem of misfearing?

The Availability Heuristic

The first problem is purely cognitive: the use of the availability heuristic in thinking about risks (Noll and Krier, 1990). It is well-established that people tend to think that events are more probable if they can recall an incident of its occurrence.[1] Consider, for example, the fact that people typically think that more words on any given page will end with the letters *ing* than have *n* as the second-to-last letter (though a moment's reflection shows that this is not possible) (Tversky and Kahneman, 1983). With respect to risks, judgments are typically affected by the availability heuristic, so that people overestimate the number of deaths from highly publicized events (motor-vehicle accidents, tornados, floods, botulism) but underestimate the number from less publicized sources (stroke, heart disease, stomach cancer) (Baron, 1994, p. 218). Similarly, much of the concern with nuclear power undoubtedly stems from its association with memorable events, including Hiroshima, Chernobyl, and Three-Mile Island.

Consider in this regard a cross-national study of perceptions of risk associated with terrorism and SARS (Feigenson, Bailis, and Klein, 2004). Americans perceived terrorism to be a far greater threat, to themselves and to others, than SARS; Canadians perceived SARS to be a greater threat, to themselves and to others, than terrorism. Americans estimated their chance of serious harm from terrorism as 8.27%, about four times as high as their estimate of their chance of serious harm from SARS (2.18%). Canadians estimated their chance of serious harm from SARS as 7.43%, significantly higher than their estimate for terrorism (6.04%). Notably, the figures for SARS were unrealistically high, especially for Canadians; the best estimate of the risk of contracting SARS, based on Canadian figures, was .0008% (and the chance of dying as a result less than .0002%). For obvious reasons, the objective risks from terrorism are much harder to calculate, but if it is estimated that the United States will suffer at least one terrorist attack each year with the same number of deaths as on September 11, the risk of death from terrorism is about .001%—a speculative number under the circumstances, but not an implausible place to start.

What accounts for the cross-national difference and for the generally exaggerated risk perceptions? The availability heuristic provides a large part of the answer. In the United States, risks of terrorism have (to say the least) received a great deal of attention, producing a continuing sense of threat. But there have been no incidents of SARS, and the media coverage has been limited to events elsewhere—producing a degree of salience, but one far lower than that associated with terrorism. In Canada, the opposite is the case. The high degree of public discussion of SARS cases, accompanied by readily available instances, produced an inflated sense of the numbers—sufficiently inflated to exceed the same numbers from terrorism (certainly a salient risk in Canada, as in most nations post–9/11).

To the extent that people lack information or base their judgments on mental shortcuts that produce errors, a highly responsive government is likely to blunder. Indeed, private groups often enlist availability, emphasizing an incident that is supposed to be taken as

representative of a much larger problem. Cost-benefit analysis is a natural corrective, above all because it focuses attention on the actual effects of regulation, including, in some cases, the existence of surprisingly small benefits from regulatory controls. To this extent, cost-benefit analysis should not be taken as undemocratic, but, on the contrary, should be seen as a means of fortifying (properly specified) democratic goals by ensuring that government decisions are responsive to well-informed public judgments.

Aggravating Social Influences: Informational and Reputational Cascades

The availability heuristic does not, of course, operate in a social vacuum. It interacts with emphatically social processes, and in particular with informational and reputational forces (Kuran and Sunstein, 1999). When one person says, through words or deeds, that something is or is not dangerous, he creates an informational externality (Caplin and Leahy, 1998). A signal by some person A will provide relevant data to others. When there is little private information, such a signal may initiate an informational cascade, with significant consequences for private and public behavior, and often with distorting effects on regulatory policy (Kuran and Sunstein, 1999, p. 720).

Imagine, for example, that A says that abandoned hazardous waste sites are dangerous, or that A initiates protest activity because such a site is located nearby. B, otherwise skeptical or in equipoise, may go along with A; C, otherwise an agnostic, may be convinced that if A and B share the relevant belief, the belief must be true; and it will take a confident D to resist the shared judgments of A, B, and C. The result of this set of influences can be social cascades, as hundreds, thousands, or millions of people come to accept a certain belief simply because of what they think other people believe (Hirschleifer, 1995). There is nothing fanciful about the idea. Cascade effects help account for the existence of widespread public concern about abandoned hazardous waste dumps (a relatively trivial environmental hazard), and they spurred grossly excessive public fears of the pesticide Alar, of risks from plane crashes, and of dangers of shootings in schools in the aftermath of the murders in Littleton, Colorado. Such effects helped produce massive dislocations in beef production in Europe in connection with "mad cow disease"; they have also spurred European fear of genetic engineering of food.

On the reputational side, cognitive effects may be amplified as well. If many people are alarmed about some risk, you may not voice your doubts about whether the alarm is merited simply not to seem obtuse, cruel, or indifferent. And if many people believe that a certain risk is trivial, you may not disagree through words or deeds, lest you appear cowardly or confused. The result of these forces can be cascade effects, mediated by the availability heuristic. Such effects can produce a public demand for regulation even though the relevant risks are trivial. At the same time, there may be little or no demand for regulation of risks that are, in fact, quite large in magnitude. Self-interested private groups can exploit these forces, often by using the availability heuristic. Consider the fact that European companies have tried to play up fears of genetically engineered food as a way of fending off American competition.

Cost-benefit analysis has a natural role here. If agencies are disciplined by that form of analysis, they will have a degree of insulation from cascade effects induced by informational and reputational forces, especially when the availability heuristic is at work. The effect of cost-benefit analysis is to subject misfearing to a kind of technocratic scrutiny, to ensure that the public demand for regulation is not rooted in myth, and to ensure as well that government is regulating risks even when the public demand (because insufficiently informed) is low. And here too there is no democratic problem with the inquiry into consequences. If people's concern is fueled by informational forces lacking much reliability, or if people express concern even though they are not fearful, a technocratic constraint on "hot" popular reactions is hardly inconsistent with democratic ideals. Similarly, there is nothing undemocratic about a governmental effort to divert resources to serious problems that have not been beneficiaries of cascade effects.

Emotions and Probability Neglect

As a result of the availability heuristic, people can have an inaccurate assessment of probability. But sometimes people venture little assessment of probability at all, especially when strong emotions are involved. In such cases, large-scale variations in probabilities will matter little—even when those variations unquestionably should matter a great deal. What affects thought and behavior is the outcome, not the likelihood that it will occur. Here too is a problem of misfearing.

The phenomenon of probability neglect received its clearest empirical confirmation in a striking study of people's willingness to pay to avoid electric shocks (Rottenstreich and Hsee, 2001). One experiment attempted to see whether varying the probability of harm would matter more, or less, in settings that trigger strong emotions than in settings that seem relatively emotion-free. In the "strong emotion" setting, participants were asked to imagine that they would participate in an experiment involving some chance of

a "short, painful, but not dangerous electric shock." In the relatively emotion-free setting, participants were told that the experiment entailed some chance of a $20 penalty. Participants were asked to say how much they would be willing to pay to avoid participating in the relevant experiment. Some participants were told that there was a 1% chance of receiving the bad outcome (either the $20 loss or the electric shock); others were told that the chance was 99%.

The central result was that the variations in probability affected those facing the relatively emotion-free injury, the $20 penalty, far more than they affected people facing the more emotionally evocative outcome of an electric shock. For the cash penalty, the difference between the median payment for a 1% chance and the median payment for a 99% chance was predictably large: $1 to avoid a 1% chance, and $18 to avoid a 99% chance. For the electric shock, by contrast, a large difference in probability made little difference to median willingness to pay: $7 to avoid a 1% chance, and $10 to avoid a 99% chance! Apparently people will pay a significant amount to avoid a small probability of an emotionally laden hazard, and the amount that they will pay will not vary greatly with the changes in probability.

There is much evidence in the same vein. Consider these findings:

1. When people discuss a low-probability risk, their concern rises even if the discussion consists mostly of apparently trustworthy assurances that the likelihood of harm is small (Alkahami and Slovic, 1994).
2. If people are asked how much they will pay for flight insurance for losses resulting from "terrorism," they will pay more than if they are asked how much they will pay for flight insurance from all causes (Loewenstein et al., 2001).
3. People show "alarmist bias." When presented with competing accounts of danger, they tend to move toward the more alarming account (Viscusi, 1997).
4. In experiments designed to test levels of anxiety in anticipation of a painful electric shock of varying intensity, the probability of the shock had no effect. "Evidently, the mere thought of receiving a shock is enough to arouse individuals, but the precise likelihood of being shocked has little impact on level of arousal" (Viscusi, 1997).

It is important to be careful with the relevant categories here (Elster, 1999; Kahan and Nussbaum, 1996; Nussbaum, 1999). Emotions are generally the products of beliefs, and hence an emotional reaction to risk—terror, for example—is generally mediated by judgments. But this is not always true; sometimes the operation of the brain ensures intense emotional reactions with minimal cognitive activity (Loewenstein et al., 2001). In any case, the judgments that fuel emotions may be unreliable. We need not venture into controversial territory in order to urge that some risks seem to produce extremely sharp, largely visceral reactions. Indeed, experience with "mass panics" has shown exactly this structure, as assurances based on statistical evidence have little effect in the face of vivid images of what might go wrong.[2]

The role of cost-benefit analysis is straightforward here. Just as the Senate was designed to have a "cooling effect" on the passions of the House of Representatives, so cost-benefit analysis might ensure that policy is driven not by hysteria or alarm, but by a full appreciation of the effects of relevant risks and their control.

Nor is cost-benefit analysis, in this setting, only a check on unwarranted regulation. It can and should serve as a spur to regulation as well. If risks do not produce visceral reactions, partly because the underlying activities do not yield vivid mental images, cost-benefit analysis can show that they nonetheless warrant regulatory control. The elimination of lead in gasoline, which was driven by cost-benefit analysis, is a case in point.

Systemic Effects and "Health-Health Tradeoffs"

Often people focus on small pieces of complex problems, and causal changes are hard to trace. The German psychologist Dietrich Dorner has conducted some illuminating computer experiments designed to see whether people can engage in successful social engineering (Dorner, 1996). Participants are asked to solve problems faced by the inhabitants of some region of the world. Through the magic of the computer, many policy initiatives are available to solve the relevant problems (improved care of cattle, childhood immunization, drilling more wells). But most of the participants produce eventual calamities, because they do not see the complex, system-wide effects of particular interventions. Only the rare participant can see a number of steps down the road—to understand the multiple effects of one-shot interventions on the system.

Often regulation has similar systemic effects. A decision to regulate nuclear power may, for example, increase the demand for coal-fired power plants, with harmful environmental consequences (Breyer, 1978; see also Huber, 1987). A decision to impose fuel economy standards on new cars may cause a "downsizing" of the fleet and in that way increase risks to life. A decision to ban asbestos may cause

manufacturers to use less safe substitutes. Regulation of ground-level ozone may control the health dangers of ozone, but ozone has various benefits as well, including protection against cataracts and skin cancer; hence regulation of ozone may cause health problems equal to those that it reduces.[3] Indeed, the regulation of ozone will increase electricity prices, and because higher electricity prices will deprive poor people of air conditioning or lead them to use it less, such regulation may literally kill people (Gray, 1998).

These are simply a few examples of situations in which a government agency is inevitably making "health-health tradeoffs" in light of the systemic effects of one-shot interventions. Indeed, any regulation that imposes high costs will, by virtue of that fact, produce some risks to life and health, since "richer is safer" (Cross, 1995; Graham, Chang, and Evans, 1992; Keeney, 1994; Wildavsky, 1980, 1988). A virtue of cost-benefit analysis is that it tends to overcome people's tendency to focus on parts of problems by requiring them to look globally at the consequences of apparently isolated actions.

Dangers On-Screen, Benefits Off-Screen

Why are people so concerned about the risks of nuclear power, when experts tend to believe that those risks are quite low—lower, in fact, than the risks from competing energy sources, such as coal-fired power plants, which produce relatively little public objection? Why do ordinary people tend to believe that the small risks from pesticides should be regulated, even if the comparatively small risks from X-rays are quite tolerable?

Suggestive answers come from research suggesting that for many activities that pose small risks but that nonetheless receive intense public concern, people perceive low benefits as well as high risks.[4] For example, many people see nuclear power as a low-benefit, high-risk activity. Similar findings appear for some activities that are in fact relatively high-risk: a judgment of "low risk" accompanies a judgment of "high benefits." The very fact that they are known to have high benefits skews judgment in their favor, and hence makes people understate the costs as well.

The obvious conclusion is that sometimes people favor regulation of some risks because the underlying activities are not seen to have compensating benefits.[5] Thus for some activities, tradeoffs are not perceived at all. The dangers are effectively on-screen, but the benefits are off-screen. Note that this is not because such activities do not, in fact, have compensating benefits. It is because of a kind of perceptual illusion, a cognitive bias.

An important factor here is loss aversion, which leads people to see a loss from the status quo as more undesirable than a gain is seen as desirable (Camerer, 1995; Kahneman, Knetsch, and Thaler, 1990; Thaler, 1991).[6] In the context of risk regulation, the consequence is that any newly introduced risk, or any aggravation of existing risks, is seen as a serious problem, even if the accompanying benefits (a gain from the status quo and hence perceived as less salient and less important) are considerable.[7] Thus when a new risk adds danger, people may focus on the danger itself and not on the benefits that accompany the danger. And an important problem here is that in many cases where the dangers are on-screen and the benefits off-screen, the magnitude of the danger is actually quite low. Cost-benefit analysis can be a corrective, by placing the various effects on-screen.

General Implications

The cognitive argument for cost-benefit analysis is now in place. It is true but obvious to say that people lack information and that their lack of information can lead to an inadequate or excessive demand for regulation, or a form of "paranoia and neglect" (Graham, 1996). What is less obvious is that predictable features of cognition will lead to a public demand for regulation that is unlikely to be based on the facts. Political entrepreneurs should be expected to exploit those features of cognition, attempting (for example) to enlist availability and probability neglect so as to create fear. And when the costs of risk reduction are on-screen, people may misfear in the sense of neglecting serious dangers. Consider the United States on September 10, 2001, or in the period preceding Hurricane Katrina; in both cases, experts argued in favor of far more concern than people were willing to show.

For purposes of sensible resource allocation, the goal should be to create procedural checks in the form of legal safeguards against an excessive or insufficient demand for regulation. When people ask for regulation because of fear fueled by availability cascades, and when the benefits from the risk-producing activity are not registering, it would be highly desirable to impose cost-benefit filters on their requests through legal requirements that must be surmounted before regulation may be imposed. When interest groups exploit cognitive mechanisms to create unwarranted fear or to diminish concern with serious problems, it is desirable to have institutional safeguards established as real constraints on excessive and insufficient action.

When people fail to ask for regulation for related reasons, it would be desirable to create a mechanism

by which government might nonetheless be encouraged to act if the consequences of action would be desirable. Here, too, cost-benefit balancing might be desirable, as in fact it has proved to be in connection not only with the phaseout of lead but also with the Reagan administration's decision to phase out CFCs, which was motivated by a cost-benefit analysis suggesting that the phaseout would do far more good than harm.[8]

A caveat: It is entirely possible that the public demand for regulation will result from something other than cognitive errors, even if the relevant risk seems low as a statistical matter. People may think, for example, that it is especially important to protect poor children from a certain risk in a geographically isolated area, and they may be willing to devote an unusually large amount to ensure that protection. What seems to be a cognitive error may turn out, on reflection, to be a judgment of value, and a judgment that can survive reflection. I will return to this point. For the moment note two simple points. Whether an error is involved is an empirical question, subject, at least in principle, to empirical testing. And nothing in cost-benefit analysis would prevent people from devoting resources to projects that they consider worthy, even if the risk is relatively low as a statistical matter.

Of course it is true that cost-benefit analysis might operate as a solution to a range of problems, not only the cognitive ones that I have been emphasizing here. If interest groups are able to obtain regulation that is in their interest and to block regulation that is in the public interest, cost-benefit balancing should serve as a safeguard. If people are myopic and treat the future as if it is valueless, cost-benefit analysis can be a significant help. Consider, for example, the problem of climate change, where misfearing has produced insufficient action and where excessive discounting of the future has also proved an obstacle to desirable controls. Cost-benefit analysis helps show that such controls are desirable, as in the context of ozone-depleting chemicals, where aggressive regulation was spurred by that form of analysis (Sunstein, 2007), and restrictions on greenhouse gases were reasonably defended, through international action, by a careful analysis of what might be lost and what might be gained.

Objections: Populism, Quantification, and Rival Rationalities

The argument made thus far, cautious though it may seem, runs into three obvious objections. The first involves democratic considerations; the second points to the limitations of quantification; and the third involves the possibility that ordinary people's judgments are based not on cognitive limitations but on a kind of "rival rationality."

Populism

The initial objection, which is populist in character, is that in a democracy, government properly responds to the social "demand" for law. Government does not legitimately reject that demand on the grounds that cost-benefit analysis suggests that it should not act. On this view, a democratic government should be accountable. Any approach that uses efficiency or technocratically driven judgments as a brake on accountability is fatally undemocratic. People may or may not misfear, but a democratic government pays careful attention to their concerns.

The problem with this objection is that it rests on a controversial and even unacceptable conception of democracy, one that sees responsiveness to citizens' demands, whatever their factual basis, as the foundation of political legitimacy. However, if those demands are uninformed, it is perfectly appropriate for government to resist them. Indeed, it is far from clear that reasonable citizens want, or would want, their government to respond to their uninformed demand. The foregoing analysis thus far suggests that the relevant demands are, in fact, uninformed or unreflective. If this is so, they should be subject to deliberative constraints of the sort exemplified by cost-benefit analysis. After that analysis has been generated and public officials have taken it into account, democratic safeguards continue to be available, and electoral sanctions can be brought to bear against those who have violated the public will. At the very least, cost-benefit analysis should be an ingredient in the analysis, showing people that the consequences of various approaches might be different from what they seem.

Qualitative Differences among Social Goods

Some people object to cost-benefit analysis on the grounds that many of the goods at stake in regulation (human and animal life and health, recreational and aesthetic opportunities) are not merely commodities, that people do not value these goods in the same way that they value cash, and that cost-benefit analysis, by virtue of its reductionism, is inconsistent with people's reflective judgments about the issues at stake. Arguments of this sort have been developed into philosophical challenges to efforts to turn various goods into what analysts deem to be monetary equivalents (Anderson, 1993).

Such arguments are convincing if cost-benefit analysis is taken to suggest a controversial position in favor

of the commensurability of all goods—if cost-benefit analysis is seen to insist that people should value environmental amenities, or their own lives, in the same way that they value a bank account, or if cost-benefit is taken as a metaphysical claim to the effect that all goods can be aligned along a single metric, or if five lives saved is seen as the same, in some deep sense, as $20 or $30 million saved. Part of what people express in their daily lives is a resistance to this form of commensurability, and some goods are understood to have intrinsic as well as instrumental value. The existence of qualitative differences among goods fortifies the claim that any "bottom line" about costs and benefits should be supplemented with a more qualitative description of the variables involved.

But cost-benefit analysis need not be seen as embodying a reductionist account of the good, and much less as a suggestion that everything is simply a "commodity" for human use. It is best taken as a pragmatic instrument, agnostic on the deep issues and designed to assist people in making complex judgments where multiple goods are involved. To put it another way, cost-benefit analysis might be assessed pragmatically, or even politically, rather than metaphysically. We should conclude that the final number may provide less information than the ingredients that went into it, and that officials should have and present cost-benefit analysis in sufficiently full terms to enable people to have a concrete sense of the effects of regulation. This is an argument against some overambitious understandings of what cost-benefit balancing entails. But it is not an argument against cost-benefit balancing.

Rival Rationalities

The final objection to the cognitive argument for cost-benefit analysis is, in a sense, the most fundamental. On this view, cost-benefit analysis is not desirable as a check on ordinary intuitions because those intuitions reflect a kind of "rival rationality." Ordinary people do not always misfear, even when they deviate from experts. On the contrary, they have a complex understanding of what it is that they want to maximize. They do not simply tabulate lives saved; they ask questions as well about whether the relevant risk is controllable, voluntary, dreaded, equitably distributed, and potentially catastrophic. Consider table 13.1.

Some people suggest that to the extent that ordinary people disagree with experts, they have a "thicker" or "richer" rationality, and that democracy should respect their judgments (Slovic, Fischhoff, and Lichtenstein, 1985). On a more moderate view, government's task is to distinguish between lay judgments that are products of factual mistakes (produced, for example, by the availability heuristic) and lay judgments that are products of judgments of value (as in the view that voluntarily incurred risks deserve less attention than those that are involuntarily incurred) (Pildes and Sunstein, 1995). In any case the "psychometric paradigm" is designed to show how ordinary people's judgments are responsive to an array of factors other than lives saved (Slovic, 1997).

The simplest response to this claim is that it need not be a criticism of cost-benefit analysis at all; it may suggest only that any judgment about benefits and

Table 13.1

Risk traits	Aggravating	Mitigating
Familiarity	New	Old
Personal control	Uncontrollable	Controllable
Voluntariness	Involuntary	Voluntary
Media attention	Heavy media coverage	Ignored by media
Equity	Unfairly distributed	Equitably distributed
Children	Children at special risk	Children not at risk
Future generations	Future generations at risk	Future generations not at risk
Reversibility	Irreversible	Reversible
Identifiability of victims	Victims known	Victims not identifiable
Accompanying benefits	Benefits clear	Benefits invisible
Source	Human origin	Created by nature
Trust in relevant institutions	Low trust in relevant institutions	High trust in relevant institutions
Timing of adverse effects	Effects delayed	Effects immediate
Understanding	Mechanisms poorly understood	Mechanisms understood
Precedents	History of accidents	No past accidents

costs (whether or not based on willingness to pay) will have to take account of people's divergent assessments of different risks. In principle, there is no problem with doing exactly that. There is, however, reason to question the now-conventional view that qualitative factors of this kind fully explain people's disagreement with experts about certain risks of death. No doubt it is *possible* that people's judgments about risk severity are a product of some of the more qualitative considerations listed above; this idea leads to the widespread view that ordinary people have a "richer" rationality than do experts, since ordinary people look at the nature and causes of death, not simply at aggregate deaths at issue. But it is also possible that an apparently "rich" judgment that a certain risk is severe, or not severe, depends not on well-considered judgments of value but on the operation of System I. More particularly, people's judgments may depend instead on an absence of ordinary contextual cues, on a failure to see that tradeoffs are inevitably being made, on heuristic devices that are not well adapted to the particular context, or instead on a range of confusing or confused ideas that people cannot fully articulate (Margolis, 1998).

When people say, for example, that the risk of nuclear power is very serious, they may be responding to their intense visceral concern, possibly based on (uninformed) statistical judgments about likely lives at risk and on their failure to see (as they do in other contexts) that that risk is accompanied by a range of social benefits. Thus it is possible that a judgment that a certain risk of death is unusually bad is not a "rich" qualitative assessment but an (unreliable) intuition based on a rapid balancing that prominently includes perceived lives at stake and the perceived presence of small or no benefits associated with the risk-producing activity. If no such "richer rationality" is involved, cost-benefit analysis can proceed as a check of misfearing. If richer rationality is in fact the source of people's judgments, then the valuation of costs and benefits should incorporate their beliefs.

An Incompletely Theorized Agreement on Cost-Benefit Analysis?

Problems With Aggregated Willingness to Pay

Thus far I have suggested that cost-benefit analysis is a sensible approach to cognitive problems faced by ordinary people in the assessment of risk. I have also suggested that there is no democratic objection to using cost-benefit analysis as an ingredient, even a crucial ingredient, in decisions and that cost-benefit analysis can be understood in a way that responds to reasonable concerns about quantification and about the idea that the only thing to be maximized is total lives saved (or, somewhat better, life-years saved).

But none of this deals with the general question about how cost-benefit analysis should be understood. In the least contentious formulation—the formulation that I have used here—cost-benefit analysis is simply a form of open-ended consequentialism, an invitation to identify the advantages and disadvantages of regulation, an invitation that does not say anything about appropriate weights. The virtue of this formulation is that it is uncontentious; the vice is that it is vacuous. People can agree with it, but it does not mean anything. In its most contentious formulation, cost-benefit analysis depends on asking people how much they are "willing to pay" for various goods and on making decisions that depend on the resulting numbers. Problems with this approach lie in a possible lack of private information, its possible distributional unfairness (since willingness to pay depends on ability to pay), potential differences between private willingness to pay and public aspirations, and collective-action problems of various sorts that might draw into doubt the privately expressed amounts.

For present purposes, the most serious difficulty is that willingness to pay may be a product of misfearing. People's judgments may be insufficiently informed or unreflective with respect to both facts and values. For example, people may overstate the risks from various risks that receive disproportionate media attention. If this is so, it seems odd to base government policy on those judgments. It is also possible that people will be willing to pay little to avoid some bad X simply because they are used to it and their preferences have adapted accordingly. Preferences based on lack of information or adaptation to deprivation are hardly a good basis for regulatory policy. They need not be taken as given and translated into law.

Incomplete Theorization: Cost-Benefit Analysis As Political, Not Metaphysical

Often it is possible to resolve hard questions of law and policy without resolving deeply contested issues about justice, democracy, or the appropriate aims of the state (Sunstein, 1996, 1999). Often it is possible to obtain an incompletely theorized agreement on a social practice, and even on the social or legal specification of the practice. In many areas of law and public policy, people can reach closure about what to do despite their disagreement or uncertainty about why, exactly, they ought to do it. Thus people who disagree about the purposes of the criminal law can agree that rape and murder should be punished, and

punished more severely than theft and trespass. Thus people can support an Endangered Species Act amid disagreement about whether the protection of endangered species is desirable for theological reasons, or because of the rights of animals, plants, and species, or because of the value of animals, plants, and species for human beings. A great advantage of incompletely theorized agreements is that they allow people of diverse views to live together on mutually advantageous terms. An even greater advantage is that they allow people of diverse views to show one another a high degree of both humility and mutual respect.

I believe that incompletely theorized agreement is possible here; at least this should be the goal of those attempting to understand the uses of cost-benefit analysis in regulatory policy. For the reasons just discussed, it would be difficult to obtain agreement on the view that all questions of regulatory policy should be resolved by asking how much people are "willing to pay" for various social goods. But it should nonetheless be possible for diverse people to agree on presumptive floors and ceilings for regulatory expenditures, and the presumptions can do a great deal of useful work for policy making and for law. In short, a great deal can be done without confronting the hardest theoretical questions raised by contentious specifications of cost-benefit analysis.

An obvious question here is: Who could join this incompletely theorized agreement? Who would reject it? My principal claim is that the agreement could be joined by a wide range of reasonable people, including utilitarians and Kantians, perfectionist and political liberals, and those who accept and those who doubt the idea that private willingness to pay is the appropriate foundation for regulatory policy. There is room here for deliberative democrats who emphasize the need for government to reflect on private preferences, rather than simply to translate them into law.[9] A prime purpose of the approach is to ensure more in the way of reflection; cost-benefit analysis, as understood here, is a guarantee of greater deliberation, not an obstacle to it. Nor is the approach rigid. Under the proposed approach, agencies have the authority to abandon the floors and ceilings if there is reason for them to do so. If, for example, agencies want to spend a great deal to protect African American children from a risk disproportionately faced by them, they are entitled to do so, as long as they explain that this is what they are doing and as long as what they are doing in reasonable.

Eight Propositions

Here, then, are eight propositions offered in the hope that they might attract support from diverse theoretical standpoints. The goal is to provide a starting point for the effort to anchor cost-benefit analysis in an incompletely theorized agreement about regulatory policies.

1. *Agencies should identify the advantages and disadvantages of proposed courses of action and also attempt to quantify the relevant effects to the extent that this is possible.* When quantification is not possible, agencies should discuss the relevant effects in qualitative terms and also specify a range of plausible outcomes, e.g., annual savings of between 150 and 300 lives, or savings of between $100 million and $300 million, depending on the rate of technological change. The statement should include the full range of beneficial effects.

2. *The quantitative description should supplement, rather than displace, a qualitative description of relevant effects.* Both qualitative and quantitative descriptions should be provided. It is important to know the nature of the relevant effects, for example, lost workdays, cancers averted, respiratory problems averted. To the extent possible, the qualitative description should give a concrete sense of who is helped and who is hurt, for example, whether the beneficiaries are mostly or partly children, whether the regulation will lead to lost jobs, higher prices, more poverty, and so forth. Where the only possible information is speculative, this should be noted, along with the most reasonable speculations.

3. *Agencies should attempt to convert nonmonetary values (involving, for example, lives saved, health gains, and aesthetic values) into dollar equivalents.* This is not because a statistical life and, say, $5 million are the same thing, but to promote coherence and uniformity and to ensure sensible priority setting. The conversion into monetary equivalents is simply a pragmatic tool to guide analysis and to allow informed comparisons.

4. *Agencies entrusted with valuing life and health should be controlled, either by statute or executive order, via presumptive floors and ceilings.* For example, a statute might say that a statistical life will ordinarily be valued at no less than $6 million and no more than $12 million. Evidence of worker and consumer behavior, suggesting a valuation of between $5 million and $7 million per statistical life saved, is at least relevant here. The fact that the willingness to pay numbers are in this range is hardly decisive, but it is supplemented by the fact that similar numbers appear to represent the midpoint of agency practice. If an agency is going to spend, say, no more than $500,000 per life saved, or more than $20 million, it should have to explain itself.

5. *Agencies should be permitted to adjust the ceilings and floors or to choose a low or high end of the range on the basis of a publicly articulated and*

reasonable judgment that such an adjustment or such a choice is desirable. Perhaps adjustments could be made if, for example, poor people are especially at risk. There should be no adjustments "downward" for poor people; in other words, the fact that poor people are willing to spend less to protect their own lives (because they are poor) should not call for correspondingly lower expenditures by government. The principal danger here is that well-organized groups will be able to use equitable arguments on behalf of their preferred adjustments. It is important to ensure a degree of discipline here, and perhaps the dangers of interest-group manipulation are serious enough to suggest that uniform numbers or ranges might be used or that the presumptions are strong and rebuttable only in the most compelling cases.

6. *Agencies should be permitted to make adjustments on the basis of the various "qualitative" factors discussed above.* For example, they might add a "pain and suffering premium" or increase the level of expenditure because children are disproportionately affected or because the victims are members of a disadvantaged group. It would be reasonable to conclude that because AIDS has disproportionate adverse effects on homosexuals and poor people, special efforts should be made to ensure against AIDS-related deaths. To the extent possible, they should be precise about the nature of, and grounds for, the relevant adjustments, especially in light of the risk that interest-group pressures will convert allegedly qualitative adjustments in illegitimate directions.

7. *The appropriate response to social fear not based on evidence and to related "ripple effects" is education and reassurance rather than increased regulation.* The best response to misfearing is educational; the government should not expend significant resources merely because an uninformed public believes that it should. But if education and reassurance fail, increased regulation may be defensible as a way of providing a kind of reassurance in the face of intense fears, which can themselves impose high costs of various kinds. (Consider, for example, the possibility that people who are afraid of the risks of plane crashes will shift to driving, a riskier method of transportation; consider also the fact that the fear is itself a cost.)

8. *Unless the law explicitly requires otherwise, judicial review of risk regulation should require a general showing that regulation has produced more good than harm, based on a reasonable view about the valuation of both benefits and costs.* On this view, courts should generally require agencies to generate and to adhere to ceilings and floors. But they should also allow agencies to depart from conventional numbers (by, for example, valuing a life at less than $6 million or more than $10 million) if and only if the agency has given a reasonable explanation of why it has done so. The ultimate task would be to develop a kind of "common law" of cost-benefit analysis, authorizing agencies to be law-making institutions in the first instance.

Conclusion

I have suggested that cost-benefit analysis, often defended on economic grounds, can be urged less contentiously on cognitive grounds. Cost-benefit analysis, taken as an inquiry into the consequences of varying approaches to regulation, is a sensible response not only to interest-group power but also to the problem of misfearing. The underlying problems include the use of the availability heuristic; social amplification of that heuristic via cascade effects; probability neglect; a misunderstanding of systemic effects, which can lead to unanticipated bad (and good) consequences; and a failure to see the benefits that accompany certain risks. In all of these areas, an effort to identify costs and benefits can helpfully inform analysis.

My ultimate hope is that it would be possible to produce a convergence on a form of cost-benefit analysis that should be understood as a pragmatic instrument and that ought not to be terribly contentious—a form of cost-benefit analysis that does not take a stand on highly controversial questions about what government ought to do and that promises to attract support from people with diverse conceptions of the right and the good. I have suggested here that the most promising source of such an agreement lies not only, or even mostly, in neoclassical economics but instead in an understanding of the general problem of misfearing.

Notes

This chapter was written before Sunstein joined the federal government, first as a senior adviser to the director of the Office of Management and Budget, and later as administrator of the Office of Information and Regulatory Affairs. It should go without saying that nothing here represents, in any way, an official position of the United States government. Thanks to Eldar Shafir for valuable comments on a previous draft.

1. Tversky and Kahneman (1982) describe the availability heuristic.

2. See the discussion of Love Canal in Karan and Sunstein (1999).

3. Lutter and Wolz (1997) estimated that the EPA's new ozone NAAQS could cause 25 to 50 more melanoma skin cancer deaths and increase the number of cataract cases by 13,000 to 28,000 each year. Keeney and Green (1997)

calculated that if attainment of the new standards costs $10 billion annually, a number well within EPA's estimated cost range, it would contribute to 2,200 premature deaths annually. On the general phenomenon, see Graham and Wiener (1995).

4. The fact that nuclear power and application of pesticides produce benefits as well as risks may not "register" on the lay viewscreen, and this may help produce a "high risk" judgment (Alkahami and Slovic, 1994).

5. See Margolis (1998) for a detailed discussion of how this point bears on the different risk judgments of experts and lay people.

6. Thaler (1991) argues that "losses loom larger than gains" (p. 143).

7. For some policy implications of loss aversion, see Knetsch (1997).

8. See Lutter and Wolz (1997) and Keeney and Green (1997). The Reagan administration supported aggressive regulation largely because cost-benefit analysis from the Council of Economic Advisers demonstrated that "despite the scientific and economic uncertainties, the monetary benefits of preventing future deaths from skin cancer far outweighed costs of CFC controls as estimated either by industry or by EPA" (Benedick, 1991, p. 63).

9. Absolutists of various kinds might refuse to join an agreement on these principles. Perhaps their refusal would be most reasonable in the case of the Endangered Species Act, where nothing said below explains why millions of dollars should be spent (at least in opportunity costs) to save members of ecologically unimportant species. It would be possible, however, to imagine a kind of "meta" cost-benefit analysis that would point in this direction, perhaps on the ground that it greatly simplifies decision making without imposing high costs overall. For the regulatory issues dealt with here, an absolutist approach seems hard to justify, not least because there are dangers to life and health on both sides of the equation.

References

Alkahami, A. S., and Slovic, P. (1994). A psychological study of the inverse relationship between perceived risk and perceived benefit. *Risk Analysis*, *14*, 1085–1096.

Anderson, E. (1993). *Value in ethics and economics.* Cambridge, MA: Harvard University Press.

Baron, J. (1994). *Thinking and deciding* (2nd ed.). Cambridge: Cambridge University Press.

Benedick, R. E. (1991). *Ozone diplomacy: New directions in safeguarding the planet.* Cambridge, MA: Harvard University Press.

Breyer, S. (1978) *Vermont Yankee* and the court's role in the nuclear energy controversy. *Harvard Law Review*, *91*, 1833–1845.

Camerer, C. (1995). Individual decision making. In J. H. Kagel and A. E. Roth (Eds.), *The handbook of experimental economics* (pp. 665–670). Princeton, NJ: Princeton University Press.

Caplin, A., and Leahy, J. (1998). Miracle on Sixth Avenue: Information externalities and search. *Economic Journal*, *108*, 60–74.

Cross, F. B. (1995). When environmental regulations kill: The role of health/health analysis. *Ecology Law Quarterly*, *22*, 729–784.

Dorner, D. (1996). *The logic of failure: Why things go wrong and what we can do to make them right.* New York: Metropolitan Books.

Elster, J. (1999). *Alchemies of the mind: Rationality and the emotions.* Cambridge: Cambridge University Press.

Feigenson, N., Bailis, D., and Klein, W. (2004). Perceptions of terrorism and disease risks: A cross-national comparison. *Missouri Law Review*, *69*, 991–1012.

Graham, J. D. (1996). Making sense of risk: An agenda for Congress. In R. W. Hahn (Ed.), *Risks, costs, and lives saved: Getting better results from regulation* (pp. 183–207). Oxford: Oxford University Press.

Graham, J.D., Chang, B.-H., and Evans, J. S. (1992). Poorer is riskier. *Risk Analysis*, *12*, 333–337.

Graham, J. D., and Wiener, J. B. (1995). *Risk versus risk: Tradeoffs in protecting health and the environment.* Cambridge, MA: Harvard University Press.

Gray, C. B. (1998). The Clean Air Act under regulatory reform. *Tulane Environmental Law Journal*, *11*, 235 .

Hirshleifer, D. (1995). The blind leading the blind: Social influence, fads, and informational cascades. In M. Tommasi and K. Ierulli (Eds.), *The new economics of human behavior* (pp. 188–215). Cambridge: Cambridge University Press.

Huber, P. (1987). Electricity and the environment: In search of regulatory authority. *Harvard Law Review*, *100*, 1002–1065.

Kahan, D. M., and Nussbaum, M. C. (1996). Two conceptions of emotion in the criminal law. *Columbia Law Review*, *96*, 269–374.

Kahneman, D., Knetsch, J. L., and Thaler, R. H. (1990). Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy*, *98*, 1325–1348.

Keeney, R. L. (1994) Mortality risks induced by the costs of regulations. *Journal of Risk and Uncertainty*, *8*, 95–110.

Keeney, R. L., and Green, K. (1997). *Estimating fatalities induced by economic impacts of EPA's ozone and particulate standards.* Reason Public Policy Institute, Policy Study No. 225. Los Angeles, CA: Reason Foundation.

Knetsch, J. L. (1997). Reference states, fairness, and choice of measure to value environmental changes. In M. H. Bazerman, D. M. Messick, A. E. Tenbrunsel, and K. A. Wade-Benzoni (Eds.) *Environment, ethics, and behavior: The psychology of environmental valuation*

and degradation (pp. 13–32). San Francisco: New Lexington Press.

Kuran, T., and Sunstein, C. R. (1999). Availability cascades and risk regulation. *Stanford Law Review*, *51*, 683–768.

Loewenstein, G., Weber, E., Hsee, C., and Welch, N. (2001). Risk as feelings. *Psychological Bulletin*, *127*, 267–286.

Lutter, R., and Wolz, C. (1997). UV-B screening by tropospheric ozone: Implications for the NAAQS. *Environmental Science and Technology*, *31*, 142A–146A.

Margolis, H. (1998). *Dealing with risk: Why the public and the experts disagree on environmental issues.* Chicago: University of Chicago Press.

Noll, R. G., and Krier, J. E. (1990). Some implications of cognitive psychology for risk regulation. *Journal of Legal Studies*, *19*, 747, 749–760.

Nussbaum, M. (1999). *Upheavals of thought: The intelligence of emotions.* Cambridge: Cambridge University Press.

Pildes, R. H., and Sunstein, C. R. (1995). Reinventing the regulatory state. *University of Chicago Law Review*, *62*, 1–129.

Rottenstreich, Y., and Hsee, C. (2001). Money, kisses, and electric shocks: On the affective psychology of risk. *Psychological Science*, *12*, 185–190.

Slovic, P. (1997). Trust, emotion, sex, politics and science: Surveying the risk assessment battlefield. *University of Chicago Legal Forum*, *44*, 59–142.

Slovic, P., Fischhoff, B., and Lichtenstein, S. (1985). Regulation of risk: A psychological perspective. In R. G. Noll (Ed.), *Regulatory policy and the social sciences* (pp. 241–283). Berkeley, CA: University of California Press.

Sunstein, C. R. (1996). *Legal reasoning and political conflict.* Oxford: Oxford University Press.

———. (1999). *One case at a time: Judicial minimalism on the Supreme Court.* Cambridge, MA: Harvard University Press.

———. (2002). *Risk and reason.* New York: Cambridge University Press.

———. (2007). *Worst-case scenarios.* Cambridge, MA: Harvard University Press.

Thaler, R. H. (1991). The psychology of choice and the assumptions of economics. In R. H. Thaler (Ed.), *Quasi rational economics* (pp. 137–166). New York: Russell Sage Foundation.

Tversky, A., and Kahneman, D. (1982). Judgment under uncertainty: Heuristics and biases. In D. Kahneman, P. Slovic, and A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 3–11). Cambridge: Cambridge University Press.

———. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.

Viscusi, W. K. (1997). Alarmist decisions with divergent risk information. *Economic Journal*, *197*, 1657–1670.

Wildavsky, A. (1980, Summer). Richer is safer. *National Affairs*, *60*, 23–39.

———. (1988). *Searching for safety.* New Brunswick, NJ: Transaction Books.

CHAPTER 14

Choice Architecture and Retirement Saving Plans

SHLOMO BENARTZI

EHUD PELEG

RICHARD H. THALER

On March 28, 1979, the Unit 2 nuclear power plant on the Three Mile Island Nuclear Generating Station in Dauphin County, Pennsylvania, suffered a core meltdown. In the investigation that followed, it became clear that a valve that was supposed to regulate the flow of cooling water had failed. The operators sent a control signal to remotely shut the valve, and when they received an indication that the signal had been sent, they assumed that the valve was indeed shut. An actual "positive feedback" lamp indicating the true position of the valve did not exist, so the operator had no way of verifying whether the signal was received and the necessary actions taken. Such a lamp was deemed expendable during the construction of the facility to save time and money. As a result of this design error, the operators were unaware that the valve was *not* turned off, that the cooling water continued to pour out, and that the reactor's core continued to overheat and eventually melted down. Even though initial reports blamed "human error," subsequent investigations found the design of controls equally at fault. They determined that ringing alarms and flashing warning lights left operators overwhelmed by information, much of it irrelevant, misleading, or incorrect.[1]

The Three Mile Island incident is one example of how a faulty design can lead even highly qualified decision makers to devastating results. Although managing a retirement portfolio is not a national mission-critical operation, a financial meltdown can be just as painful to an individual as a plant meltdown is to the masses. In this article, we propose that the design of nuclear-plant control rooms, everyday objects, and retirement saving vehicles share similar properties.

There are two crucial factors to consider. First, everything matters. Tiny details, from the color of an alert lamp to the size of the font can influence choices. Second, since everything matters, it is important for those who design choice environments whom Thaler and Sunstein (2008) call "choice architects," to take human factors into account. Choice architecture is particularly important in domains such as retirement savings, where most of the decision makers are unsophisticated.

Prior research in the domain of retirement savings has illustrated the potential role of improved choice architecture. Madrian and Shea (2001), for example, showed that the choice of default has a dramatic effect on savings behavior. They studied several plans that changed the default so that employees who took no action were automatically enrolled into the retirement savings plan. It is important to note, however, that freedom of choice was preserved because employees could always opt out of the retirement plan and were not in any way forced to save. In one of the plans studied, the percentage of employees saving for retirement increased from 49% to 86% when the default was changed to automatically enrolling employees into the plan.

Other studies have also documented that design does matter. Benartzi and Thaler (2001), for example, showed that the menu of investment funds offered to employees affects their risk-taking behavior. In particular, some employees follow a naive diversification strategy of spreading their money equally across funds, something they have dubbed the 1/n rule. As a result of using the 1/n rule (or a variant of this rule), a plan offering a bond fund, a small cap stock fund, and a large cap stock fund might result in employees leaning toward an allocation of two-thirds in stocks. In comparison, a plan with a money market fund, a bond fund, and a diversified stock fund might result in just one-third allocated to stocks.[2]

Iyengar and Kamenica (2006) documented that the size of the menu of funds also affects savings behavior. They studied a cross-section of retirement savings

plans, some offering as few as 2 funds and others with as many as 59 funds. They estimated that the addition of 10 funds to the menu of choices decreased participation in the plan by 2%, because some employees might have been overwhelmed by the degree of choice.

The intuitive principle that many minor design elements could end up being important also applies to retirement saving vehicles. Benartzi and Thaler (2007), for example, showed that the number of lines displayed on the investment election form could have the unintended consequence of influencing the number of funds people choose. In one experiment they conducted, visitors to the Morningstar.com website (an online provider of financial information) were presented with an investment election form that had either four or eight lines displayed. (Note that those who were presented with four lines could still select more than four funds by simply clicking on a link to the form with eight lines.) Benartzi and Thaler found that only 10% of those presented with four lines ended up picking more than four funds versus 40% for those who saw eight lines on their form to begin with. In other words, the graphic designer who creates the investment election form could accidentally influence the number of funds people pick.

In this paper, we provide new evidence on choice architecture in the domain of retirement savings plans and will focus on two timely design issues related to the Pension Protection Act of 2006 (hereafter, PPA).[3] The first design issue has to do with escalator programs, where employees precommit to periodic saving increases (see Thaler and Benartzi, 2004). Whereas PPA encourages the use of escalator programs, there are many design elements that are left to the discretion of the employer, such as the timing of the saving increases. Since every design element could end up being important, we explored the effect of a variety of design issues in this context. Our goal was to identify the choice architecture that helps employees save more. As Choi et al. (2002) reported, most employees (68%) feel they are saving too little, so we were just trying to identify the choice architecture that helps people reach their own stated goal.

Consistent with the work of Madrian and Shea (2001), we found that inertia plays a crucial role in choice architecture. In particular, we found that when the escalator program was set as an opt-in program, about 15%–25% of new hires signed up for the program. In contrast, when employees were automatically enrolled in the escalator program, only 16.5% opted out and the remaining 83.5% ended up in the escalator program.[4] We also found that seemingly minor design elements do matter in the context of escalator programs. For example, we document that employees prefer to precommit to save more next January as opposed to, say, next February or next March. In the spirit of New Year's resolutions, people seem to think that January is a good time to start exerting willpower.

The second design issue we explored has to do with portfolio solutions. In recent years, fund providers have come up with one-stop portfolio solutions to assist employees with the complicated task of fund selection. One solution offered by fund providers is risk-based funds. These funds are often labeled conservative, moderate, or aggressive, and employees are expected to pick the one fund that matches their risk preferences. A distinctive feature of risk-based funds is that they keep a constant asset allocation and do not reduce their equity exposure as people get older. A competing solution offered by fund providers is retirement date funds. These funds are often labeled 2010, 2020, 2030, and 2040, where the labels correspond to the expected retirement date. Unlike risk-based funds, retirement date funds decrease their equity exposure as people approach retirement. In the case of retirement date funds, employees who are looking for a simple portfolio solution should pick the fund that matches their expected retirement date.

One might view the packaging of bond funds and stock funds into one-stop portfolio solutions as inconsequential, since individuals still have access to the underlying bond funds and stock funds to select the mix of funds they truly prefer. However, we found that one-stop portfolio solutions increased equity market participation by about three percentage points. More importantly, the effect was larger for lower-income employees, hence it reduced the well-documented gap in equity market participation between lower-income and higher-income individuals. We also found that retirement date funds strengthened the negative correlation between age and risk-taking behavior. It is important to note that the stronger negative correlation between age and risk taking was observed not only for investors in retirement date funds, but also for the entire population of participants in plans offering retirement date funds. Understanding how the architecture of one-stop portfolio solutions affects investor behavior is essential in light of the PPA and the related guidelines by the Department of Labor that bless a spectrum of one-stop portfolio solutions.

In the next section, we will discuss choice architecture and the effectiveness of escalator programs in increasing saving rates, and in the following section, we will then consider the effects of choice architecture on portfolio choices. In both discussions, we will provide new evidence that design matters and that seemingly minor design elements can end up being important. We will provide concluding remarks in the last section.

Choice Architecture and Escalator Programs

Background

The worldwide trend toward defined contribution retirement plans has shifted the responsibility for retirement planning from the employer to employees. In most defined contribution plans, employees must determine how much to save for retirement and how to invest their funds. Given the difficulty of calculating the "optimal" saving rate as well as the presence of self-control problems, it should not come as a surprise that most people are not saving enough to maintain a comfortable lifestyle at retirement (Skinner, 2007). And as we noted earlier, 68% of plan participants agreed that their saving rate was too low (Choi et al., 2002).

Being interested in helping people reach their stated goal of saving more, we used the basic psychological principles of hyperbolic discounting, inertia, and nominal loss aversion to design a program that helps employees increase their saving rates. The program offers individuals the opportunity to precommit to automatic saving increases, which could take place every time someone receives a pay raise, or alternatively, on a set date, such as every January 1. Of course, participants in the program can always change their mind and either stop the automatic saving increases or quit saving altogether. We dubbed the program Save More Tomorrow (hereafter, SMarT).[5]

Features of SMarT were incorporated into the PPA, which encourages employers to automatically enroll new and existing employees into their retirement savings plans. The act prescribes an initial saving rate of at least 3% of pay, an annual increase increment of at least 1%, and a target rate of at least 6%, but no more than 10%. Employers who follow the above guidelines and provide a generous matching contribution are exempt from the nondiscrimination tests (i.e., they do not have to prove that lower-paid employees are benefiting fairly from the retirement plan in comparison to higher-paid employees). Note that the act allows for saving increases to take place on any date and does not require that saving increases and pay raises be synchronized. Similar legislative initiatives are taking place in the U. K. and New Zealand.[6]

Retirement-plan providers also expressed great interest in automatic saving increases. Vanguard made the program available to more than one million employees, Fidelity Investments (2006) reported that 6,000 of its employer clients had offered the program to their employees, and T. Rowe Price and TIAA-CREF, among other providers, also rolled out similar programs. A survey by Hewitt Associates (2007) indicated that 31% of plan sponsors had offered the program to their employees in 2006, and that 42% of those who had not were likely to do so in 2007. Similar programs were also introduced in the U. K. and Australia. The rapid penetration of the program into the marketplace reflects the importance of choice architecture.

The accumulating data on the program suggest dramatic cross-sectional differences in employee take-up rates. In our original case study with one-on-one financial counseling, take-up rates reached 80% (Thaler and Benartzi, 2004). With automatic enrollment into the program, participation rates also reached 80%. On the other end of the spectrum, some retirement service providers reported take-up rates as low as a few percentage points.

In this section, we will attempt to identify the design elements of the program that are most effective at helping people better reach their retirement savings goals. Our research was driven by both theoretical and practical interests. From a theoretical perspective, we were interested in better understanding the psychology of saving. From a practical perspective, we were interested in fine-tuning the program to help more people save more.

We will next discuss the psychological principles underlying the program in more detail—that is, hyperbolic discounting, inertia, and nominal loss aversion. As we describe each psychological principle and design element, we will also investigate its role in the success of the program. We will also compare each design element to the specific plan design features prescribed by the PPA.

Hyperbolic Discounting

The first psychological principle that guided us in the design of the program was hyperbolic discounting, which refers to a discount function that "overvalues the more proximate satisfaction relative to the more distant ones" (Strotz, 1955, p. 177). Read and van Leeuwen (1998), for example, asked subjects to choose between healthy snacks (bananas) and unhealthy snacks (chocolate). When asked one week in advance, only 26% of the subjects indicated they would choose the unhealthy snack. However, when asked immediately prior to consuming the snack, 70% chose the unhealthy snack. This form of present-biased preferences is characterized by the discount rate increasing as consumption gets closer (see Frederick, Loewenstein and O'Donoghue, 2002; Loewenstein and Elster, 1992; and Thaler, 1981 for additional evidence).

Hyperbolic discounting and present-biased preferences could explain why many of us engage in suboptimal behavior such as excessive eating, lack of exercise,

and excessive spending. Yet, at the same time, many of us envision we will eat less, exercise more, and save more in the not-too-distant future. Hyperbolic agents believe (often wrongly) that doing the right things will be easier in the future, because the temptation to, say, eat too much will be moderated (see Laibson, 1997; and O'Donoghue and Rabin, 1999, 2001, which model such behavior). To help hyperbolic agents save more, our program invited employees to sign up to save more in the *future*. However, we did not know the role of this specific feature in the success of the program, nor did we know what the time lag should be between signing up for the program and the effective date of the first saving increase to encourage maximum participation in the program.

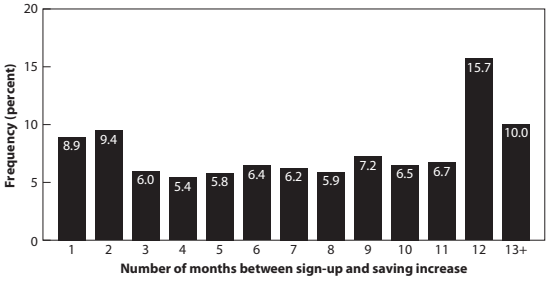
The first SMarT case study (Thaler and Benartzi, 2004) offers some insight into the role of hyperbolic discounting in the success of the program. In that case study, they found that 78% of those who declined to increase their saving rates right away agreed to do so every time they got a pay raise. This pattern of behavior is consistent with hyperbolic discount functions. However, this evidence is more a joint test of both hyperbolic discounting and nominal loss aversion, as the distant saving increases were synchronized with pay raises and employees never saw their take-home pay decrease. Here, we provide more direct evidence on the role of hyperbolic discounting.

We explored the role of hyperbolic discounting in several ways. First, we obtained data from Vanguard, a large provider of retirement plan services that rolled out an automatic increase service called OneStep Save at the beginning of 2004.[7] We looked at 65,452 plan participants in 273 plans who were hired when these plans already offered the opportunity to join the OneStep program. Joining the program had to be done via the web or the phone at the employee's initiative. The results show that 15.1% of the new employees joined the program. Participation rates varied by plan, with an interquartile range between 4% and 19%.

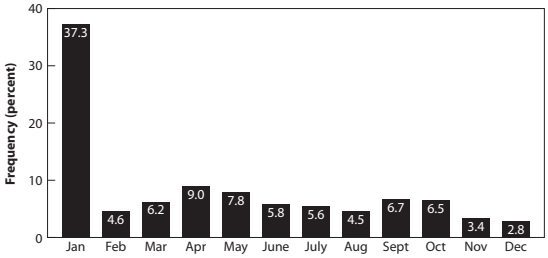
What makes the Vanguard data of particular interest for our analysis is the fact that almost all individuals had to select the timing of saving increases on their own.[8] While the saving increases take place once a year, the participant still had to select the specific month for the increase to apply as well as the saving increment. Hyperbolic agents are predicted to prefer to increase their saving rate sometime in the future, although theory does not tell us how much of a delay between joining the program and increasing savings individuals would like to have.

Figure 14.1 displays the number of months that passed between participants signing up for the program and their desired date of saving increase. The figure is based on 49,433 participants who joined the program as of year-end 2005, and it reveals some interesting differences across individuals. Some preferred to implement the program sooner rather than later. Specifically, 8.9% of participants preferred the saving increase to be implemented within the same month they signed up. However, the remaining 91.1% of participants preferred to postpone saving increases, which is consistent with hyperbolic discounting. At the extreme, 15.7% preferred that the first increase take place exactly one year after signing up, and 10.0% of participants wanted to wait longer than a year.

We suspect there could also be a time-of-year effect. January might be a good candidate, since hyperbolic agents might consider doing the right things "next year." Figure 14.2 shows the month of increase selected by the program participants. Almost 40% of the participants actually selected January as the month of increase, and no other month seemed to have such a dominating effect. (The distribution across months



14.1. Distribution of the delays between signing up for the escalator program and the effective date of the saving increase. Data was obtained from Vanguard, and it consists of 49,433 program participants who joined the program as of year end 2005. For example, 15.7% of participants requested that their first saving increase take place 12 months after joining the program.



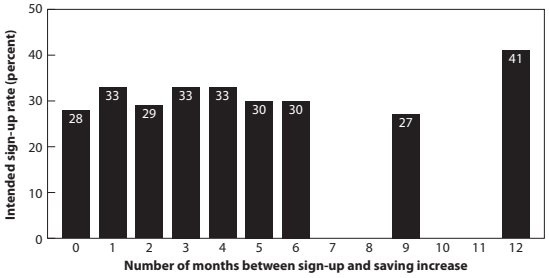
14.2. The month in which participants requested their saving rate to increase. Data was obtained from Vanguard, and it consists of 49,433 program participants. For example, 37.3% of participants requested that their saving rate go up in January.

is statistically different from a uniform distribution at the 0.01 level.)

While the Vanguard data is consistent with hyperbolic discounting, it does not measure the strength of preferences. For example, would those who postponed their saving increase by one year still join a program that was set to increase saving much earlier? To answer this question, we used data from T. Rowe Price, another large provider of retirement plan services. T. Rowe Price conducted an online survey of plan participants at a large firm during September 2005. Participants were given a short paragraph describing the program and then asked whether or not they would be interested in signing up for the program. The saving increases were set to take place in X months, where X was varied from implementing the increase immediately to postponing implementation by 12 months. Each participant responded only to one of the conditions.

Figure 14.3 displays the intended sign-up rates for the different conditions. Generally, about 30% of the participants intended to sign up. However, there is something special about postponing saving increases by 12 months, where the sign up rate was 41% ($p > 0.05$). This result is consistent with the Vanguard data, in which delaying the increase by exactly one year was more popular than any other choice.

The results are consistent with a combination of hyperbolic discounting and some type of mental accounting.[9] Models of hyperbolic agents could explain why many participants prefer to postpone saving increases, but it is not obvious why postponing the increases by 3 months, 6 months, and 9 months is equally attractive, yet postponing by 12 months is more attractive. Similarly, it is not clear why postponing to January is more attractive than any other month. We speculate that it might have something to do with the tradition of turning over a new leaf at the start of the year.



14.3. Intended sign-up rates and the delay between signing up for the program and the effective date of saving increase from an experiment conducted online by T. Rowe Price. Subjects were given a short description of a program offering automatic saving increases and were asked whether or not they would like to join. The length of time between joining the program and experiencing the saving increase was varied, though each subject was presented with one scenario only. Seven hundred and forty nine subjects participated in the survey.

The PPA provides flexibility with respect to the timing of saving increases. Hence, employers could, for example, pick January as the month of implementing saving increases to encourage employee participation. More generally, minor design elements such as the month of the saving increases could end up influencing employee saving behavior.

Inertia

The second psychological principle that guided us in the design of the program was inertia, or what Samuelson and Zeckhauser (1988) dubbed the *status-quo bias.* Inertia is known to have a dramatic effect on participants' behavior in defined contribution plans. Typically, inertia prevents individuals from taking the right actions. For example, many participants do not rebalance their portfolios at all (Samuelson and Zeckhauser, 1988), whereas others do not even join the retirement plan, even when it is a virtual arbitrage opportunity (Choi, Laibson, and Madrian, 2004). On the other hand, inertia can also be used in a positive way to enhance plan participation. For example, flipping the default so that employees are automatically enrolled in the plan, unless they take an action to opt out, increases participation rates dramatically (Madrian and Shea, 2001).[10]

The SMarT program attempts to use inertia in a positive way to help people reach their stated goals of saving more. In particular, once an individual signs up for the Save More Tomorrow program, future saving increases take place automatically unless the individual changes her mind and opts out. Another plan design option is to automatically enroll employees into the Save More Tomorrow program. So, unless someone opts out, she will automatically be in the program, and future saving increases will take place automatically as well. Here we explore the effect of automatically enrolling people into our program. The powerful evidence on the role of inertia leads us to hypothesize that default choices have a significant impact on participation rates in the program.

The first implementation of the program on an opt-out basis took place in 2003 by the Safelite Group, a client of Strong Retirement Services. The program was introduced to employees in June 2003 with an effective saving increase date of July 2003, an annual increment of 1% of pay, and no synchronization between pay raises and saving increases. It is

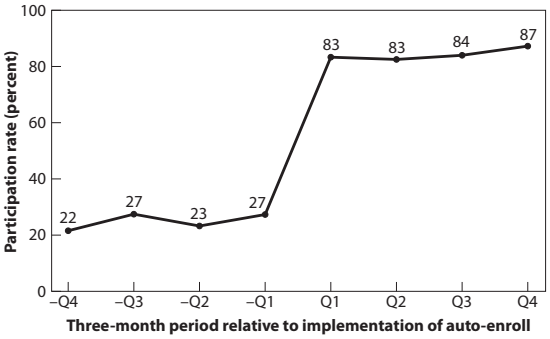
important to note that hyperbolic discounting probably did not play a major role in this setting because saving increases took place relatively soon after enrolling in the program. And nominal loss aversion should not have played a role at all in this setting since saving increases took place on a set date regardless of pay raises. Hence, this was a unique opportunity to identify and focus on the role of inertia in the success of the program.

We have summary statistics on 3,640 employees who were already participating in the 401(k) plan as of May 2003, the month prior to the introduction of the automatic increase program. Ninety-three percent of participants took no action, thus they were automatically enrolled in the program. Six percent actively opted out of the program, and the remaining one percent of participants used this opportunity to increase their saving rate beyond the automatic increase.

Since the Safelite Group implementation in July 2003, additional implementations on an opt-out basis have taken place. In our Vanguard dataset, 13 plans introduced an opt-out version of the program, one in July 2004 and the rest in 2005.[11] The opt-out programs covered new hires only, and they were typically set with an initial deferral rate of about 3% of pay and an annual increment of 1% of pay. There was substantial variation in the "cap," with some plans stopping the increases at 5% and others stopping it at 25%, and even 50%. There was also substantial variation in the default portfolio choices, with some plans selecting a money market fund and others selecting a balanced fund or retirement date funds. Hence, this opt-out version of the program was more of an autopilot 401(k) plan where enrollment, deferral rates, and portfolio choices were all automatically selected on behalf of employees, who had the option to opt out.

Figure 14.4 displays the percentage of Vanguard plan participants who took part in the contribution escalator before and after the introduction of automatic enrollment, and it is based on 2,222 new employees who were eligible for the contribution escalator when they were hired. In the 12 months prior to the implementation of automatic enrollment, 25.1% opted into the contribution escalator. However, in the 12 months following automatic enrollment, 83.5% of the savers were participating in the escalator program. (The differences were statistically significant at the 0.01 level.) The dramatic change in participation illustrates the power of inertia and the important role of choice architecture.[12]

One caveat is that the opt-out program was generally introduced in 2005 with the first saving increase scheduled for 2006. Hence, we could not determine from our data, which ends with 2005, how many participants, if any, opted out right before the increase.



14.4. Employee participation rates before and after the implementation of auto-enroll automatic contribution escalator for 13 retirement saving plans that introduced automatic enrollment of new employees into the plan as well as into a contribution escalator. The study was based on 2,222 plan participants.

Data from the one plan that introduced the program in 2004 and already had the first increase in 2005 suggest an opt-out rate of just 9%, so it does not look like participants opted out right before the first increase.

Another potential caveat is that opt-out programs might "trick" employees into a program they do not really want. Choi et al. (2005) provided some insightful evidence on this issue from two sets of experiments: one having to do with automatically enrolling people into a 401(k) plan at a modest saving rate (although without automatic increases), and the other having to do with requiring employees to make an active choice, whether it was to join or not to join the plan. They found that active decision making results in participation rates that are similar to those of automatic enrollment, so it does not look like automatic enrollment tricks people into the plan (at least in their context of a modest, yet constant, saving rate). It is also important to note that there is no way to avoid setting a default, and it is not clear that the current default of having procrastinators keep their low saving rates is a better one.

Nominal Loss Aversion

The third psychological principle that guided the design of our program was nominal loss aversion. Loss aversion refers to the fact that the pain associated with losses is about twice the pleasure that is associated with similar magnitude gains (Tversky and Kahneman, 1992). To the extent that individuals view increased savings and the respective reduction in spending as a loss, loss aversion predicts that it could be difficult to help people save more. However, the

crucial factor for our program is that people tend to evaluate losses relative to some nominal reference point. For example, in a study of perceptions of fairness (Kahneman, Knetch, and Thaler, 1986), subjects were asked to judge the fairness of pay cuts and pay increases. One group of subjects was told that there was no inflation and was asked whether a 7% wage cut was "fair." A majority, 62%, judged the action to be unfair. Another group was told that there was a 12% inflation rate and was asked to judge the perceived fairness of a 5% raise. Here, only 22% thought the action was unfair.[13] Of course, in real terms the two conditions are virtually identical, but in nominal terms they are quite different.

To ensure that saving increases are not perceived as losses, our program suggests that pay raises and saving increases be synchronized. For example, the program can be designed so that every time an employee receives a pay raise, he takes home half the raise and the remaining half is contributed to the retirement plan. This design feature ensures that the take-home amount does not decrease. It is, unfortunately, easier said than done, due to some practical implementation issues. For example, very often information on pay raises is received at the last minute and there is not enough time to update the contribution rate. Hence, we were interested in understanding the role of nominal loss aversion from both a theoretical perspective and a practical perspective. To the extent that nominal loss aversion does not play an important role in the success of the program, plan sponsors and plan providers could offer employees a simplified program where saving increases take place on a set date regardless of pay raises.

Ultimately, we would like to conduct a field experiment in which employees are randomly assigned to one of two conditions. In one condition, employees would be offered the original version of the program, where saving increases and pay raises are synchronized, whereas in the other condition, saving increases would take place on a set date regardless of pay raises. There are several practical obstacles that make it very difficult to run such an experiment. First, employers are often reluctant to offer different retirement plan features to different employees due to legal concerns. Second, it is tricky to synchronize saving increases and pay raises because employees would like to get advance notice of the forthcoming saving increase, but information on pay raises is often provided at the last minute.

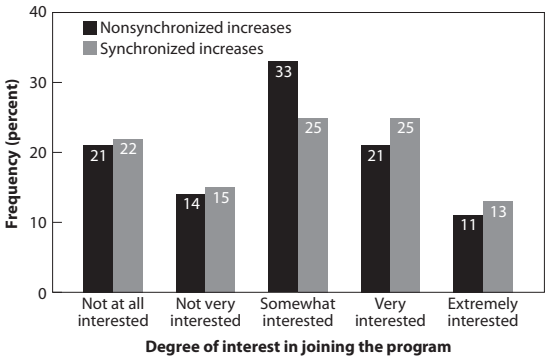
Given the above-mentioned difficulties of conducting a randomized field experiment, we decided to conduct a survey with the help of Warren Cormier of the Boston Research Group (Cormier, 2006). The survey group included 5,246 retirement-plan participants served by half a dozen different vendors. The subjects were interviewed by phone and asked for their interest in joining an automatic saving increase program. One group of subjects were told that saving increases would take place every January and there was no mention of pay raises. Specifically, they were told:

> Some 401(k) plans offer a new program to make it easy for employees to save more. If you join the program, each January the percentage of your pay that you're contributing to your plan will automatically increase by 1%, until you reach a savings rate of 15%. So if you are currently contributing 5%, the program would increase your contribution to 6%. Of course, you are in control and can stop the increases at any time.

Another group of subjects were further told that the saving increases could be synchronized with pay raises. Specifically, they were told:

> You could also choose to have the amount you're contributing automatically increase by 1% every time you get a pay raise instead of every January. With this feature, your savings will never cause your take-home pay to go down.

The results of the survey are displayed in figure 14.5. Thirty-two percent of the subjects said that



14.5. The effect of synchronizing saving increases and pay raises on the degree of interest in joining the program. The data are from an experiment conducted via telephone by the Boston Research Group, where 5,246 retirement plan participants were asked for their level of interest in joining the program. One group of participants was told that saving increases would take place every January (i.e., nonsynchronized increases), whereas another group was told that the saving increases could also take place every time they get a pay raise, so that their take-home pay would not go down (i.e., synchronized increases).

they were either very interested or extremely interested in the nonsynchronized program that automatically increased their saving rates every January regardless of pay raises. In comparison, 38% of the subjects said they were either very interested or extremely interested in the synchronized version of the program that allowed for the increases to take place every time a pay raise was received. (The difference in the degree of interest in the program is statistically significant at the 0.01 level.)

To summarize our findings so far, it seems as though inertia plays the most dominant role in the program, where defaulting employees into the program results in nearly universal participation. This result might not be as trivial and as expected as it might seem. While defaults are very powerful, employees do not always stick to the default. Anecdotal evidence from the United States indicates that a lot of employees opt out of their defined benefit plans and select the lump-sum option. Similarly, Alessandro Previtero shared with us interesting data from Italy, where more than 80% of employees opted out of the default investment for their severance package.

Hyperbolic discounting also plays a role in the success of escalator programs, with a 12-month delay between sign up and saving increases raising projected participation by roughly 10%. As to the role of nominal loss aversion, synchronizing saving increases and pay raises increased the percentage of subjects who were either very or extremely interested in joining the program by 6%. It does seem, however, that the role of nominal loss aversion is a second-order effect.

The PPA encourages automatic enrollment with an initial deferral rate of at least 3% of pay. The PPA also encourages automatic increases to a minimum deferral rate of 6% of pay. Employers who follow the prescribed guidelines are exempt from the nondiscrimination tests (i.e., they do not have to document that lower-paid employees and higher-paid employees are all benefiting from the plan).

The PPA seems to have incorporated the right design elements. It encourages automatic enrollment and automatic saving increases in line with the research findings on the powerful role of inertia in participants' behavior. In addition, the PPA provides flexibility on the timing of increases, but it remains silent on the issue of synchronization. The PPA prescribes annual saving increases, but there is no requirement that the increases be synchronized with pay raises. Given the second-order effect of nominal loss aversion in the program and the practical difficulties in synchronizing pay raises and saving increases, mandating synchronization could have been excessively burdensome.[14]

Choice Architecture and Portfolio Choices

Background

Research on participants' behavior in retirement saving plans indicates that individuals have a hard time saving enough and constructing a well-diversified portfolio (Benartzi and Thaler, 2007). Retirement-plan providers have attempted to help employees make better portfolio choices by offering simple one-stop solutions. There are at least two types of portfolio solutions in the marketplace, one being *risk-based funds* (often called lifestyle funds) and the other being *retirement date funds* (often called lifecycle funds). Risk-based funds maintain a constant level of risk, and they are often labeled conservative, moderate, or aggressive to convey their level of risk. Employees who are offered risk-based funds should simply pick the fund that best fits their risk preferences, although we must admit that figuring out your risk preference is easier said than done.

Retirement date funds are different from risk-based funds in that they follow lifecycle investment models rather than a fixed asset allocation. In particular, retirement date funds decrease their risk level as the retirement date approaches. One strategy available to employees who are offered retirement date funds is to simply pick the fund that matches their projected retirement date.

We refer to risk-based funds and retirement date funds as *asset allocation funds* (although the term we use should not be confused with tactical asset allocation funds that periodically make bets on certain asset classes). Asset allocation funds have played an increasingly important role in defined contribution plans. Hewitt Associates (2007) surveyed 146 employers and found that 57% offered retirement date funds and 38% offered risk-based funds. Policy makers also expressed interest in asset allocation funds. The Department of Labor has recently issued proposed guidelines on appropriate investments for defined contribution plans in the context of employees who are automatically enrolled into a retirement plan and are defaulted into an investment or portfolio set by their plan sponsor. The guidelines encourage the use of asset allocation funds.

Given the increasing role of asset allocation funds in retirement plans, we were interested in exploring the effect of choice architecture in this domain. In particular, we wanted to learn how the packaging of cash, bond funds, and stock funds into these one-stop portfolio solutions affects behavior. Since employees can still select any mix of cash, bonds, and stocks by using the underlying investment funds to

self-construct their own portfolio, one might view such repackaging as inconsequential. However, our data suggest that repackaging and choice architecture do matter. We will begin by describing our data and some descriptive statistics on the usage of asset allocation funds and then analyze the effect of asset allocation funds on both equity market participation and the lifecycle pattern of investing.

Data and Descriptive Statistics

Our dataset included about 1.5 million participants in 1,830 defined contribution plans served by Vanguard.[15] The data provided a snapshot of investment elections made by the participants as of December 2005. In particular, we knew the total contributions made during December 2005, the amount invested in each of the asset allocation funds, and the percentage of contributions allocated to equities, bonds, and cash. The data also included the following information for most participants: age, gender, plan entry date, account balance, and registration for the www.vanguard.com website, as well as proxies for household income and household financial wealth based on the participant's nine-digit zip code.[16] At the plan level, our data included indicators for the following plan features: the availability of loans, the inclusion of company stock in the menu of funds, and access to a brokerage account.

The Vanguard set of risk-based funds included four LifeStrategy funds, and the set of retirement date funds included six Target Retirement funds. In most cases, the sets were offered in their entirety in a given plan. The funds were classified by Vanguard under one category, called lifecycle funds, and were marketed on their website (www.vanguard.com) as "a transparent, simple-to-use solution for identifying and maintaining a proper asset allocation." Furthermore, according to the website, "the funds are designed as an investment choice for novice investors. They are the 'one-stop shopping' choice offering complete diversification in a single fund."

One of the issues we had with the data was that the information on the menu of funds in the plan was recorded as of June 2005, whereas the individual investment elections were recorded as of December 2005. To resolve this issue, we decided to determine the type of funds included in the plan by analyzing the contributions made in December 2005. We considered a plan to have offered a certain fund if at least one participant made a contribution to that fund. Based on this classification, we found that 520 plans qualified as offering retirement date funds; 811 qualified as offering risk-based funds; and 95 plans offered both types of funds.[17]

We followed Benartzi and Thaler (2001, 2002) and examined the allocations of contributions rather than the allocations of accumulated account balances. Whereas finance theory focuses on the allocation of account balances or total wealth, we preferred to study the allocation of flows into the plan. The reason for this choice was that the allocation of account balances was affected by the investment elections the participants made many years ago when they joined the plan and by subsequent fund performance. As mentioned earlier, few participants rebalanced their portfolio allocations. Another issue to consider is that retirement date funds were a relatively recent addition to the menu of funds available to plan participants. Hence, we focused on the allocations made by participants who had joined in the last two years and whom we dubbed *new participants.* Participants joining in those years were more likely to have been offered retirement date funds when they made their "critical" first selection. We also noted that retirement date funds became available in plans mostly during 2004 and 2005, whereas risk-based funds had been offered for a longer period. Some of the new participants in plans that offered retirement date funds did not have the possibility of investing in them when they joined. That is not the case for plans that offered risk-based funds.

We began our analysis by exploring who used asset allocation funds and how they were used. We found that 37% of the plan participants who were offered retirement date funds used them. However, the use of risk-based funds was somewhat higher, with 48% of those who were offered risk-based funds using them. We suspect that the lower usage of retirement date funds was related to them being newer.[18]

We also examined the cross-sectional variation in adoption rates and found that women, younger employees, and those with lower monthly contributions, income, and wealth were more likely to use retirement date funds (Table 14.1).

In particular, women were 6.0% more likely to use retirement date funds than men; employees in their 20s were 8.4% more likely to use these funds than employees in their 60s; and, employees with the lowest contributions were 6.2% more likely to use these funds than those with the highest contributions. Similar patterns emerged in the adoption of risk-based funds.

The descriptive statistics suggest that retirement date funds cater to demographic groups who are less knowledgeable about investing. For example, Dwyer, Gilkeson, and List (2002) and Lusardi and Mitchell (2005) documented that women were less knowledgeable about financial matters than men, and Kotlikoff

Table 14.1 Usage of risk-based and retirement date funds

	Risk-based funds		Retirement date funds	
Population subgroup	Hold funds (%)	100% in funds	Hold funds (%)	100% in funds
All	48.2	52.9	36.8	50.7
Gender				
Female	49.6	53.2	32.9	38.9
Male	42.5	39.0	26.9	39.7
Difference	–7.1***	–14.2***	–6.0***	0.8
Age				
20–29	50.1	57.7	40.1	46.7
30–39	48.6	49.7	36.1	45.5
40–49	47.1	50.3	34.8	57.1
50–59	46.2	54.1	34.7	61.9
60–79	43.4	55.8	31.7	67.4
Difference (old – young)	–6.7***	–1.9	–8.4***	20.7***
Monthly contributions				
0–$100	53.9	74.4	40.4	67.3
$100–$200	45.1	54.5	38.0	54.9
$200–$400	47.9	46.1	35.4	46.6
$400–$800	48.2	41.6	35.8	41.2
$800+	44.7	39.0	34.2	38.5
Difference (high – low)	–9.2***	–35.4***	–6.2***	–28.8***
Wealth				
<$5,000	47.5	56.8	36.5	58.1
$5,000–$25,000	49.8	54.8	37.9	51.1
$25,000–$50,000	49.9	52.6	37.6	48.7
$50,000–$100,000	48.6	50.6	37.4	46.7
$100,000+	45.7	48.8	35.4	45.5
Difference (high – low)	–1.8***	–8.0***	–1.1*	–12.6***
Web Registration				
Web registered	46.7	35.8	40.2	36.5
Not registered	49.6	68.2	32.6	73.1
Difference	–2.9***	–32.4***	7.6***	–36.6***

Note: The sample includes new participants in plans that offer either risk-based or retirement date funds. The sample size is 128,540 participants for plans offering risk-based funds and 74,503 participants for plans offering retirement date funds. Plans that offer both are excluded. "Hold funds" is the percentage of plan participants that include the funds in their portfolio, whereas "100% in funds" is the percentage of fund holders who hold only asset allocation funds. "Contributions" are the monthly totals for December 2005, and "Wealth" is based on the participant's zip code.
*significance at 10%
***significance at 1%

and Bernheim (2001) found a positive correlation between income and financial literacy. According to Vanguard's website, retirement date funds were "designed as an investment choice for novice investors," and our data suggest that they do serve this purpose.

In terms of the way in which asset allocation funds are being used, we investigated which employees tended to use them exclusively as a one-stop solution. We found the same demographic groups—that is, women and those with lower account balances or monthly contributions—were more likely to use them exclusively. While it is perfectly sensible for novice investors to use asset allocation funds as a one-stop solution, one wonders why the seemingly more sophisticated men and wealthier employees were not using them as the one-stop solution they were designed to be. We checked whether more sophisticated investors adopted a "core plus" strategy, where they invest most of their funds in an asset allocation fund but then have a small tilt toward a more targeted investment, such as an international fund. We found little of this type of behavior. In particular, just half (53%) of the investors in asset allocation funds invested all their contributions in these funds. Of the remaining 47%, four out of five investors placed less than half of their contributions in asset allocation funds, precluding the notion that asset allocation funds serve as the building block in a "core plus" strategy. We speculate that investors might fear investing in just one fund, not realizing that asset allocation funds are in fact well-diversified blends of several different funds.

Asset Allocation Funds and Equity Market Participation

In this section, we will analyze the effect of offering asset allocation funds to plan participants on their exposure to equity markets. As long as the equity risk premium is positive and there are no transaction costs, theory predicts that investors would own at least a small amount of equities for diversification purposes. However, researchers like Mankiw and Zeldes (1991) and Ameriks and Zeldes (2004) showed that this is not the case. Many U.S. households have no exposure to equities at all. In particular, Vissing-Jorgensen (2003) reported a large difference in participation rates between low- and high-net-worth households. By her definitions, just 18% of the former group participated in the equity markets, whereas 93% of the latter group owned stocks directly or indirectly. In this section, we will look at the effect of asset allocation funds on equity market participation.

Table 14.2 displays the fraction of plan participants who owned equities in their retirement account. We show the results for plans that (1) do not offer asset allocation funds, (2) offer risk-based funds, and (3) offer retirement date funds. Several patterns emerged from the results. First, we observed a positive correlation between various measures of wealth and equity market participation, which is consistent with earlier studies in this area. Second, asset allocation funds, be they risk-based funds or retirement date funds, increased equity market participation among those with lower income and account balances. Third, asset allocation funds did not affect equity market participation among the wealthiest. Since asset allocation funds increased equity market participation for lower-income individuals only, these funds tended to close the gap in stock ownership between lower- and higher-income participants. Specifically, we found that asset allocation funds cut the "participation gap" in approximately half. This was true whether we sorted individuals on their contributions, account balances, or income. For example, the participation gap between those with the lowest and highest plan balances was 20.8% for plans not offering asset allocation funds. That gap, however, decreased to 9.4% and 8.7% for plans offering risk-based funds and retirement date funds, respectively.

We further ran a probit regression to explain equity market participation with participant and plan attributes. The regression model is given in equation 14.1.

$$Equity_{i,j} = \alpha + \beta * [Contributions_{i,j} \mid HasAA_j] + \varepsilon_{i,j} \quad (14.1)$$

$Equity_{i,j}$ is an indicator for whether or not individual i in plan j holds any equity in his portfolio; *Contributions* is the log of the participant's total contributions in December 2005; and $HasAA_j$ is an indicator for whether or not plan j offers any type of asset allocation funds. The parameter estimates are displayed in table 14.3 with errors clustered at the plan level following Wooldridge (2003).

The results confirm that participation increased with contributions. Calculations not reported in the table indicate that doubling the monthly contributions to the plan increased the likelihood of the participant owning stocks by 5.2%. More interestingly, asset allocation funds increased equity market participation by 3.1%. And the relationship between contributions and equity market participation was diminished for plans with asset allocation funds, as indicated from the significantly negative coefficient on the interaction term between asset allocation funds and contribution level. This latter result is consistent with the univariate analysis showing that asset allocation funds raised participation in equity markets among lower-income individuals, hence closing the gap in equity market participation between low and high contributors to the plan.

Table 14.2 Equity participation gap

Population subgroup	No asset allocation funds offered	Plan offers risk-based funds	Plan offers retirement date funds
Monthly contributions			
0–$100	67.8	81.5	75.9
$100–$200	77.9	81.1	85.4
$200–$400	87.2	89.8	88.7
$400–$800	91.9	92.8	92.5
$800+	93.9	93.9	93.2
Participation gap (high – low)	26.1	12.4	17.3
Account balance			
0–$1,000	71.7	82.8	82.7
$1,000–$2,500	78.2	83.0	83.9
$2,500–$5,000	83.9	86.4	87.6
$5,000–$10,000	89.1	90.1	88.5
$10,000+	92.5	92.2	91.4
Participation gap (high – low)	20.8	9.4	8.7
Household income			
<$30,000	78.0	83.0	86.0
$30,000–$50,000	83.3	84.8	87.7
$50,000–$75,000	85.5	87.8	89.9
$75,000–$125,000	88.4	89.8	91.4
$125,000+	90.7	90.6	91.7
Participation gap (high – low)	12.7	7.6	5.7
Age			
20–29	82.9	86.0	87.5
30–39	85.6	89.0	88.0
40–49	84.0	87.7	86.5
50–59	81.5	85.3	84.7
60–79	76.4	81.7	78.9
Participation gap (young – old)	6.5	4.3	8.6

Note: The table displays the percentage of plan participants that invest in equities. We report equity market participation for participants in plans that (a) offer neither risk-based nor retirement date funds (n = 97,227), (b) offer risk-based funds (n = 119,917), and (c) offer retirement date funds (n = 69,579). *Participation gap* is the difference in equity market participation.

Why did the inclusion of asset allocation funds in the plan's menu affect equity market participation? Moreover, why did it increase participation among lower-paid employees? One reason could be that these funds reduced participation costs in the equity market, either in terms of fees or by reducing the psychic costs of choosing a fund. In the case of Vanguard, there was no difference in fees since the Vanguard retirement date funds charged the same fees as the underlying funds they owned. We thus favor the view that the presence of these funds reduced psychic costs.

Other research also supports the psychic costs argument. Charles Schwab, for example, highlighted the time-saving argument on their website (www

Table 14.3 Equity market participation

	(1)		(2)	
Variable	Coefficient	Standard error	Coefficient	Standard error
HasAA	1.023***	0.377	1.084***	0.329
Contributions	0.367***	0.032	0.322***	0.032
HasAA*Contributions	–0.161***	0.058	–0.173***	0.050
Constant	–0.994	0.213	–0.880***	0.351
Controls	–		+	
Participants, plans	329,024	1,772	328,192	1,744

Note: The table provides regression results for the following probit model:

$$Equity_{i,j} = \alpha + \beta^{*}[Contributions_{i,j} \mid HasAA_{j}] + \varepsilon_{i,j}$$

The dependent variable is an indicator equal to 1 if the participant invests in equities. The regressors are the log of the monthly contributions in December 2005 and an indicator equal to 1 if the plan offers any type of asset allocation funds. Column (1) presents regression results without plan level controls, and column (2) presents results with the following plan-level controls: portion of female participants, average contributions, average account balance, average tenure, percentage of web-registered users, whether the plan offers a loan, company stock or brokerage account, and the size of the plan as proxied by the log number of participants. Errors are clustered at the plan level to further account for plan-level effects.

.schwab.com) by asking "Are you looking for a way to reach your retirement goals, but do not have the time to actively manage your portfolio?" Vissing-Jorgensen (2003) used the psychic costs argument to explain the participation gap between individuals with low and high account balances. In her model, there was a fixed cost of learning about equity investments, measured as *X* number of hours. Since wealthier individuals could earn more dollars from participating in the equity market, they could afford the fixed costs of learning about stocks. We agree that wealthier individuals could earn more dollars from participating in equity markets, but it also costs them more to spend *X* hours learning about stocks, since they earn a higher hourly wage. As a result, it is not obvious that wealthier individuals have more of an incentive to participate in equity markets than lower-paid individuals. And therefore, it is unlikely that the implicit costs argument drove our results.

Another explanation for equity market nonparticipation was offered by Barberis, Huang, and Thaler (2006). They suggested that *narrow framing*, the tendency to evaluate the components of one's portfolio rather than the overall portfolio, could magnify the risk of investing in stocks. In our setting, the narrow framing hypothesis would imply that some participants were wary of holding equity funds even when they constructed a well-diversified portfolio because they were focused on and were averse to experiencing losses in any element of their portfolio. Asset allocation funds could mitigate narrow framing by making the individual components of the portfolio less "accessible." Note, however, that asset allocation funds could mitigate narrow framing consciously or unconsciously. For example, investors might be aware that asset allocation funds invest in equities but find it palatable since the volatility of stock returns is not segregated. Alternatively, investors might not even be aware that asset allocation funds invest in equities.

Retirement Date Funds and Lifecycle Investment Patterns

Despite extensive theoretical work on the relation between investment horizon and optimal risk-taking behavior, academic "prescriptions" are still mixed on whether or not there should be a relation between age and portfolio choices as well as on the exact form of the relation. Seminal work by Samuelson (1969) and Merton (1969) suggested that under certain conditions, the optimal allocation to the risky asset should remain constant over the life cycle. In other words, portfolio choices should be independent of both age and wealth. On the other hand, Bodie, Merton, and Samuelson (1992) and Viceira (2001) incorporated labor income and human capital as part of one's overall portfolio and came to a different conclusion. In particular, they proposed that the allocation to the risky asset should decrease with age. Most financial advisors agree with this advice. One often quoted rule of thumb is that a person's asset allocation to equities should be equal to 100 minus her age.[19]

The empirical evidence on actual lifecycle investment patterns is also mixed. Bodie and Crane (1997) found a strong negative relation between age and the fraction of the portfolio invested in stocks. Holden and Derhei (2005) also found a negative relation between equity holdings and age in a large sample of 401(k) plans. Ameriks and Zeldes (2004) used the different research approach of separating the decision

to own any stocks from the decision of how much stock to own. They find that older people are less likely to own stocks, a result driven mainly by plan participants either selling all of their equity holdings at retirement or annuitizing. However, conditional on owning some stock, they found very little correlation between age and the fraction invested in stocks.

Our main interest was whether the inclusion of asset allocation funds in the plan alters the relation between age and risk-taking behavior. Although it is not theoretically clear what the "correct" relation should be, we were able to show that the menu of funds presented to individuals affected employees' choices. Panel A of table 14.4 compares the average equity exposure of new participants who held risk-based funds to those who held retirement date funds in December 2005. We broke the two samples down by age groups as follows: 20–29, 30–39, 40–49, 50–59, and 60–79. Participants who held both types of funds were excluded.

The relation between age and equity holdings was relatively flat for those investing in risk-based funds. In particular, it went down from about 69% in stocks for those in their 20s and 30s to 62% in stocks for those in their 60s. In contrast, investors in retirement date funds exhibited a much stronger correlation between age and risk-taking behavior. In particular, the allocation to stocks decreased from 80% for the youngest group to 43% for the oldest group. We tested the differences in equity holdings between those who owned risk-based versus retirement date funds using both an ANOVA test and a Mann-Whitney Wilcoxon rank test. And we confirmed that investors in retirement date funds hold significantly more equity at the beginning of the life cycle and significantly less equity at later stages of the life cycle.

We also compared the two groups to a benchmark group of individuals in plans that offered neither type of funds. We found that investors in risk-based funds displayed a risk-age relationship that was close to the benchmark, whereas retirement-date investors had an average of 11% more in stocks in their 20s and 16% less in stocks when they were over 60.

The fact that investors who held retirement date funds displayed a strong correlation between age and equity holdings was not expected by construction, since investors could have utilized other funds to achieve their desired lifecycle pattern of risk taking. However, the more interesting finding is that investors who used risk-based funds did not appear to have a pronounced relation between age and risk-taking behavior. Note that we focused our analysis on new hires, so inertia could not explain the observed pattern.

One caveat is that there could be a selection bias in who chooses asset allocation funds. For example, it is plausible that those who choose retirement date funds prefer to decrease their portfolio risk as they get older, whereas those who choose risk-based funds have a preference for a constant allocation. Furthermore, it is plausible that investors in asset allocation funds would have picked the exact same risk level even if they had not had access to asset allocation funds and had to self-construct their portfolios.

To avoid the selection bias discussed above, we considered the effect of having access to asset allocation funds for all the participants in the plan and not just those selecting asset allocation funds. Panel B of table 14.4 displays the results. Again, there was a stronger downward-sloping relation between age and risk-taking behavior for plans offering retirement date funds than plans offering risk-based funds or plans offering neither.

In Panel C of table 14.4, we eliminated participants who did not own any equities from the analysis for two reasons. First, similar to Ameriks and Zeldes (2004), we attempted to separate the decision to have any stock from the choice of how much stock to own. Second, all asset allocation funds in our sample invested in equities, so it seemed consistent to compare investors in asset allocation funds to the population of participants investing in stocks. We observe similar patterns for this subsample. Specifically, the average equity exposure in plans offering risk-based funds was 74% for the youngest participants and slightly lower, 68%, for the oldest group of participants. Again, plans offering retirement date funds exhibited a stronger correlation between age and equity exposure, with the youngest participants having 80% in stocks versus 61% for the oldest group.

Another caveat is that the menu of funds available to employees could reflect their underlying preferences. For example, one might argue that plan administrators who select retirement date funds had previously realized that participants in their plans are inclined to have portfolios that become more conservative with age. We addressed this potential bias by looking at the lifecycle pattern of investing for participants in plans offering asset allocation funds who decided to self-construct their portfolios. Specifically, we examined those who held at least two funds, none of which was an asset allocation fund. Requiring a minimum of two funds increased the likelihood of the participant self-constructing his portfolio as opposed to being defaulted into a fund by the employer. We found a relatively flat relation between age and equity exposure for plans offering risk-based funds and retirement date funds (remember, the plans offer asset allocation funds, but our analysis focuses on those not picking the asset allocation funds). This result suggests that employees in plans offering risk-based versus retirement date funds are unlikely to be dramatically different a priori.

Table 14.4 Average allocation to equity by age

Panel A – Participants investing in asset allocation funds					
	Investors in risk-based		Investors in retirement date		Benchmark
Age	Equity (%)	Difference from benchmark	Equity (%)	Difference from benchmark	Equity (%)
20–29	69.4	+0.5***	79.6	+10.7***	68.9
30–39	69.9	–0.4	73.7	+3.7***	70.0
40–49	68.5	+0.3*	62.8	–5.4***	68.2
50–59	65.0	+0.3	53.0	–11.7***	64.7
60–79	61.8	+2.6***	43.0	–16.2***	59.2
Young – old	7.6***		36.6***		9.7***
Panel B – All participants in plans that offer asset allocation funds					
	Plans offer risk-based funds		Plans offer retirement date funds		Benchmark
Age	Equity (%)	Difference from benchmark	Equity (%)	Difference from benchmark	Equity (%)
20–29	64.0	+0.8***	69.7	+6.5***	63.2
30–39	67.0	+0.5**	68.1	+1.6***	66.5
40–49	64.7	+0.7***	62.4	–1.6***	64.0
50–59	60.2	+0.7**	56.3	–3.2***	59.5
60–79	55.8	+2.7***	48.3	–4.8***	53.1
Young – old	8.2***		21.4***		10.1***
Panel C – Participants in plans that offer asset allocation funds, conditional on owning equity					
	Plans offer risk-based funds		Plans offer retirement date funds		Benchmark
Age	Equity (%)	Difference from benchmark	Equity (%)	Difference from benchmark	Equity (%)
20–29	74.4	–1.8***	79.7	+3.5***	76.2
30–39	75.3	–2.3***	77.4	–0.2	77.6
40–49	73.8	–2.4***	72.2	–4.0***	76.2
50–59	70.6	–2.4***	66.4	–6.6***	73.0
60–79	68.4	–1.0	61.1	–8.3***	69.4
Young – old	6.0***		18.6***		6.8***

Note: This table displays the percentage of the portfolio invested in equities. Panel A includes participants that invest in asset allocation funds, though those investing in *both* risk-based and retirement date funds are excluded. Panel B includes all participants in plans that offer either risk-based or retirement date funds, but not both. Panel C is restricted to participants in these plans who have some exposure to equities, so participants without any equity exposure are excluded. The benchmark column refers to the average fraction allocated to equities in plans that do not offer asset allocation funds.

*averages are significantly different at 10%

**averages are significantly different at 5%

***averages are significantly different at 1%

We further accounted for the plan-selection bias using a regression model. Table 14.5 shows the results of a censored regression (the lower bound is 0% and the upper bound is 100%) for the percentage of equity in the portfolio against age, whether the plan had risk-based or retirement date funds, and interaction terms as specified in equation 14.2. Regression results are reported in column (1) without plan-level controls and in column (2) with plan-level controls. The plan-level controls included portion of female participants, average monthly contribution, average account balance, average tenure, percentage of web-registered users, whether the plan offered loans, company stock, or a brokerage account, and the size of the

Table 14.5 Effect of risk-based and retirement date funds on equity allocation

	(1)		(2)	
Variable	Coefficient	Standard error	Coefficient	Standard error
Age	2.031***	0.429	1.852***	0.468
Age^2	–0.028***	0.005	–0.028***	0.006
HasRB	3.824	11.885	5.981	9.864
HasRB*Age	–0.175	0.550	–0.309	0.454
HasRB*Age^2	0.003	0.006	0.004	0.006
HasRD	31.886***	8.532	25.953***	8.582
HasRD*Age	–1.228***	0.416	–1.009***	0.391
HasRD*Age^2	0.011***	0.005	0.009**	0.005
Intercept	30.473***	8.937	35.945***	12.252
Controls	–		+	
N Observations	329,024		328,192	

Note: The table provides regression results for the following censored regression:

$$PctEquity_{i,j} = \beta^*[Age_{i,j}\ HasRB_j|Age_{i,j}\ HasRD_j|Age_{i,j}] + \varepsilon_{i,j}$$

$PctEquity_{i,j}$ is the percentage invested in equities by participant *i* in plan *j*. $HasRB_j$ and $HasRD_j$ are indicators for whether plan *j* offers risk-based funds and retirement date funds, respectively. The regression model is estimated as a censored regression (lower bound = 0, upper bound = 100%). Column 1 presents regression results without plan level controls, and column 2 presents results with the following plan-level controls: portion of female participants, average contributions, average account balance, average tenure, percentage of web-registered users, whether the plan offers a loan, company stock or brokerage account, and the size of the plan as proxied by the log number of participants. Errors are clustered at the plan level to further account for plan-level effects.

**coefficient is different than zero at 5%.

*** coefficient is different than zero at 10%

plan using the log number of participants as a proxy. Errors are clustered at the plan level to further account for plan-level effects.

$$PctEquity_{i,j} = \alpha + \beta^*[Age_{i,j}\ HasRB| Age_{i,j}\ HasRD_j|Age_{i,j}] + \varepsilon_{i,j} \qquad (14.2)$$

When the plan did not include asset allocation funds ($HasRB = HasRD = 0$), the results indicated a hump-shaped relation between age and equity exposure, with the maximum at about age 37. The coefficients for plans that offer risk-based funds ($HasRB = 1$) were small and barely significant, indicating that the risk-based funds did not alter substantially the relation between age and risk taking. However, retirement date funds changed the fitted relationship by making it downward-sloping for ages 25 and above. The slope was also steeper at older ages, as indicated by the negative interaction coefficient.

We derived the marginal effects by calculating the expected change in allocation to equity when subtracting or adding 10 years of age from the sample average, which was about 38. Compared to the allocation at age 38, the allocation at age 28 was lower by 1.5% in plans with no asset allocation funds and lower by 1.6% in plans with risk-based funds, but it was higher by 1.8% in plans with retirement date funds. Thus, the relationship was downward sloping in early ages only when retirement date funds were offered. On the other hand, the allocation was always downward sloping between ages 38 and 48. The slope was rather flat in the former cases, 2.4% and 2.0%, and was steeper for plans with retirement date funds, 4.2%.

The last caveat we addressed was that retirement date funds were introduced throughout 2004 and 2005, and that some of these replaced risk-based funds with participants being "mapped" from the risk-based funds into the retirement date funds based on their age. It is plausible that our results were affected by inertia; that is, participants who were mapped to a retirement date fund and never bothered to change their portfolio allocations. To eliminate the possibility that our results may have been partially driven by participant inertia, we excluded 52 plans that shifted from risk-based to retirement date funds. The results were virtually identical to those reported earlier.[20]

To summarize, choice architecture does affect portfolio choices. The seemingly inconsequential packaging of cash, bonds, and stocks into one-stop asset

allocation solutions does affect investor behavior. In particular, asset allocation funds enhance equity market participation among lower-paid employees, and as a result, they reduce the equity market participation gap between lower- and higher-paid employees. We also find that the type of asset allocation funds being offered, be they risk-based funds or retirement date funds, affects the lifecycle pattern of investing.

Summary and Conclusions

We have highlighted the importance of design features of retirement plans and have argued that design does matter and seemingly inconsequential design elements could be important. We believe that the PPA is an example of good choice architecture. The main design feature of the PPA has to do with design for errors or inaction; that is, what happens if people do nothing? In the case of the PPA, employees who take no action might still save for retirement as long as their employer follows the PPA prescription of automatically enrolling employees into the plan and escalating their deferral rates periodically.

While the PPA has made great use of good choice architecture, it is important to note that there are many domains where choice architecture could be improved. Consider, for example, the Medicare Prescription Drug Program, often referred to as Medicare Part D. There are dozens of different plans offered in each state, making the decision very complicated. The plans actually differ by state, making it impossible for individuals to consult with friends or family members living in other states. There is no spell-checker, despite the difficulty of properly spelling the names of some prescription drugs. And there is no default, unless there are two eligible individuals, in which case they are assigned *randomly* to one of the plans. Part D is just one of many domains where more research on choice architecture could benefit society.

Notes

Benartzi is grateful for financial support from Reish Luftman McDaniel and Reicher and Vanguard. We are also grateful to Warren Cormier from the Boston Research Group, Jodi DiCenzo, Liat Hadar, Steve Utkus of Vanguard and Carol Waddell of T. Rowe Price for all the data and support they have provided us over the years. We are thankful to Robert Shiller, Emir Kamenica, Avanidhar Subrahmanyam and Mark Grinblatt for helpful comments.

1. A concise description of the event can be found in the U.S. Nuclear Regulatory Fact Sheet. Retrieved from http://www.nrc.gov/reading-rm/doc-collections/fact-sheets/3mile-isle.html

2. Huberman and Jiang (2006) extended the analysis used by Benartzi and Thaler and showed that employees were more likely to use the 1/n heuristic when the number of funds was small and when 100% is divisible by n. For example, only 5% of those selecting 9 funds used an approximately equal allocation across the 9 funds, whereas 53% of those using 10 funds used an equal allocation.

3. It is beyond the scope of this paper to cover all the design issues of the Pension Protection Act.

4. One legitimate concern is whether we have tricked people into the program, an issue we will address later.

5. Save More Tomorrow™ is a registered trademark of Shlomo Benartzi and Richard H. Thaler. The program is also referred to as the SMarT program, "auto-increase," and "contribution escalation." Firms are more than welcome to use the program free of charge as long as they are willing to share data for research purposes.

6. For a summary of the legislative changes around the world, see Retirement Security Project, (2006); and Iwry, (2006).

7. OneStep Save™ is a registered trademark of Vanguard.

8. One could potentially argue that requiring individuals to choose the month of saving increase is inconsistent with the spirit of the program, which is to make saving decisions as simple as possible. Choi et al. (2005) provided evidence that simplifying the enrollment process, so that individuals joining a 401(k) plan should only have to check the "yes" box to a predetermined combination of saving rate and investment elections, increases participation rates.

9. See work by Thaler (1999) on mental accounting.

10. Also see work by Johnson and Goldstein (2003) on the effect of defaults on organ donations. They find that countries with explicit consent have about 10% to 20% of people make their organs available for donation, whereas countries with implicit consent have about 90% of people make their organs available (i.e., only 10% opt out).

11. About 50 Vanguard clients are in the process of implementing the program on an opt-out basis.

12. We attempted to explore demographic differences between those who opted out and those who did not, but unfortunately, we had very little demographic information on the new hires.

13. See also work by Shafir, Diamond, and Tversky (1997).

14. Other design elements we explored are the annual increase increment and the "cap" (i.e., the rate at which saving increases stop). We found that employees are insensitive to the annual increment being 1% or 2% of pay. Similarly, we found that employees are insensitive to the cap being set at either 10% or 20% of pay. However, setting an unrealistically high cap tends to demotivate employees and reduce sign-up rates.

15. Yamaguchi et al. (2007) explored a closely related dataset and found similar results.

16. A company called IXI collects retail and IRA asset data from most of the large financial services companies. IXI aggregates the data from all companies at the nine-digit zip code level and then calculates the average household assets by zip code. On average, there are 10 to 12 households in a nine-digit zip code area. Next, IXI assigns a wealth rank (from 1 to 24) to each area. We narrow the ranks into 5 groups, with the respective ranges displayed in Table 14.1.

17. Unfortunately, the limited data we had about the plans in our sample did not allow us to determine if adoption of such funds are related to certain plan characteristics.

18. One concern is that some plans might have offered asset allocation funds as the default investment option, and we know this could have had a strong effect on take-up rates. We believe our results are unlikely to be affected by this issue for a couple of reasons. First, we would have expected far higher take-up rates had asset allocation funds been used as defaults. Second, most plan sponsors did not use asset allocation funds as defaults *prior to* the PPA.

19. Bodie and Crane (1997) described this rule of thumb and other generally accepted lifecycle investment prescriptions.

20. We attempted to use time-series data on the 52 plans that switched from risk-based to retirement date funds. Unfortunately, we did not have a sufficiently large number of new hires in those plans to conduct a meaningful analysis.

References

Ameriks, J., and Zeldes, S. P. (2004). *How do household portfolio shares vary with age?* Working paper. Columbia University.

Barberis, N., Huang, M., and Thaler, R. H. (2006). Individual preferences, monetary gambles, and stock market participation: A case for narrow framing. *American Economic Review*, *96*, 1069–1090.

Benartzi, S., and Thaler, R. H. (2001). Naive diversification strategies in retirement saving plans. *American Economic Review*, *91*(1), 79–98.

———. (2002). How much is investor autonomy worth? *Journal of Finance*, *57*, 1593–1616.

———. (2007). Heuristics and biases in retirement savings behavior. *Journal of Economic Perspectives*, *21*, 81–104.

Bodie, Z., and Crane, D. B. (1997). Personal investing: advice, theory and evidence. *Financial Analysts Journal*, *53*(6), 13–23.

Bodie, Z., Merton, R. C., and Samuelson, W. F. (1992). Labor supply flexibility and portfolio choice in a life cycle model. *Journal of Economic Dynamics and Control*, *16*, 427–449.

Choi, J. J., Laibson, D., and Madrian, B. (2004). *$100 bills on the sidewalk: Violation of no-arbitrage in 401(k) accounts.* Working paper. University of Pennsylvania.

———. (2005). *Reducing the complexity costs of 401(k) participation: The case of Quick Enrollment™.* Working paper. University of Pennsylvania.

Choi, J. J., Laibson, D., Madrian, B. C., and Metrick, A. (2002). Defined contribution pensions: Plan rules, participant decisions, and the path of least resistance. In J. Poterba (Ed.), *Tax policy and the economy* (Vol. 16, pp. 67–113). Cambridge, MA: MIT Press.

———. (2004). For better or for worse: Default effects and 401(k) savings behavior. In D. Wise (Ed.), *Perspectives in the economics of aging* (pp. 81–121). Chicago: University of Chicago Press.

———. (2005). *Optimal defaults and active decisions.* NBER Working Paper No. 11074. National Bureau of Economic Research.

Cormier, W. (2006). *BRG 2006 401(k) participant satisfaction study.* Boston Research Group, Boston, MA.

Dwyer, P. D., Gilkeson, J. H., and List, J. A. (2002). Gender differences in revealed risk taking: Evidence from mutual fund investors. *Economics Letter*, *76*, 151–158.

Fidelity Investments. (2006). *Building futures* (Vol. 6).

Frederick, S., Loewenstein, G., and O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature*, *40*, 351–401.

Hewitt Associates. (2007). *Survey findings: Hot topics in retirement: 2007.* Lincolnshire, IL: Hewitt Associates LLC.

Holden, S., and Derhei, J. V. (2005). Mutual funds and the U.S. retirement market in 2004. *Fundamentals*, *14*(4), 1–8.

Huberman, G., and Jiang, W. (2006). Offering versus choice in 401(k) plans: Equity exposure and number of funds. *Journal of Finance*, *61*, 763–801.

Iwry, J. M. (2006). *Automating saving: Making retirement saving easier in the United States, the United Kingdom and New Zealand.* RSP Policy Brief No. 2006-2. Retrieved from http://www.pewtrusts.org/uploadedFiles/wwwpewtrustsorg/Reports/Retirement_security/RSPPolicyBrief0606.pdf

Iyengar, S., and Kamenica, E. (2010). Choice proliferation, simplicity seeking, and asset allocation. *Journal of Public Economics*, *94*, 530–539.

Johnson, E. J., and Goldstein, D. (2003). Do defaults save lives? *Science*, *302*, 1338–1339.

Kahneman, D., Knetsch, J. L., and Thaler, R. H. (1986). Fairness as a constraint on profit seeking. *American Economics Review*, *76*, 728–741.

Kotlikoff, L. J., and Bernheim, B. D. (2001). Household financial planning and financial literacy: The need for new tools. In L. J. Kotlikoff (Ed.), *Essays on saving, bequests, altruism, and life-cycle planning* (pp. 427–478). Cambridge, MA: MIT Press.

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics*, *112*, 443–477.

Loewenstein, G., and Elster, J. (1992). *Choice over time.* New York: Sage.

Lusardi, A., and Mitchell, O. S. (2005). *Financial literacy and planning: Implications for retirement well-being.* Working paper. University of Michigan.

Madrian, B., and Shea, D. F. (2001). The power of suggestion: Inertia in 401(k) participation and savings behavior. *Quarterly Journal of Economics*, *116*, 1149–1525.

Mankiw, N., and Zeldes, S. (1991). The consumption of stockholders and nonstockholders. *Journal of Financial Economics*, *29*(1), 97–112.

Merton, R. C. (1969). Lifetime portfolio selection under uncertainty: The continuous time case. *Review of Economics and Statistics*, *51*(3), 247–257.

O'Donoghue, T., and Rabin, M. (1999). Doing it now or later. *American Economic Review*, *89*, 103–124.

———. (2001). Choice and procrastination. *Quarterly Journal of Economics*, *116*, 121–160.

Read, D., and van Leeuwen, B. (1998). Predicting hunger: The effects of appetite and delay on choice. *Organizational Behavior and Human Decision Processes*, *76*(2), 189–205.

Retirement Security Project. (2006). Analysis of the Pension Protection Act of 2006: Increasing participation through the automatic 401(k) and saver's credit. Retrieved from http://www.brookings.edu/~/media/Files/Projects/retirementsecurity/Fact%20Sheets/08_pension_bill_scorehand.pdf

Samuelson, P. A. (1969). Lifetime portfolio selection by dynamic stochastic programming. *Review of Economics and Statistics*, *51*(3), 239–246.

Samuelson, W., and Zeckhauser, R. J. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, 1, 7–59.

Save More for Retirement Act of 2005, S. Res. 875, 109th Congress, 1st Session (2005).

Shafir, E., Diamond, P., and Tversky, A. (1997). Money illusion. *Quarterly Journal of Economics*, *112*, 341–374.

Skinner, J. (2007). Are you sure you're saving enough for retirement? *Journal of Economic Perspectives*, *21*, 59–80.

Strotz, R. H. (1955). Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies*, *23*, 165–180.

Thaler, R. H. (1981). Some empirical evidence on dynamic inconsistency. *Economics Letters*, *8*(3), 201–207.

———. (1999). Mental accounting matters. *Journal of Behavioral Decision Making*, *12*, 183–206.

Thaler, R. H., and Benartzi, S. (2004). Save More Tomorrow: Using behavioral economics to increase employee savings. *Journal of Political Economy*, *112*(1), Part 2, S164–S187.

Thaler, R. H., and Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness.* New Haven, CT: Yale University Press.

Tversky, A., and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–232.

Viceira, L. M. (2001). Optimal portfolio choice for long-horizon investors with nontradable labor income. *Journal of Finance*, *56*, 433–470.

Vissing-Jorgensen, A. (2003). Perspectives on behavioral finance: Does "irrationality" disappear with wealth? Evidence from expectations and actions. *NBER Macroeconomics Annual*, *18*, 139–208.

Wooldridge, J. M. (2003). Cluster-sample methods in applied econometrics. *American Economic Review*, *93*(2), 133–138.

Yamaguchi, T., Olivia, S. M., Mottola, G. R., and Utkus, S. (2007). *Lifecycle funds in 401(k) plans.* Working paper. Pension Research Council.

CHAPTER 15

Behavioral Economics Analysis of Employment Law

CHRISTINE JOLLS

The employment relationship is often one of life's most important relationships. In both the United States and other countries, this relationship is subject to a wide range of legal requirements. Some of these legal rules regulate the formation and conduct of labor unions, while other rules govern employer-employee relationships regardless of employees' union status. The present essay discusses some of the central ways in which the second set of rules—often referred to in the United States as "employment law"—may be analyzed using behavioral economics. Because both the effects and the normative desirability of employment law turn in significant part on how employees and employers act in response to this law, behavioral economics holds clear importance for studying employment law.

This essay will draw upon a typology of behavioral economics offered by Thaler (1996). According to Thaler, behavioral economics can be characterized in terms of three "bounds" on human behavior: bounded willpower—people have trouble conforming their actions to their previously made plans; bounded self-interest—people depart from neoclassical economic notions of material self-interest maximization; and bounded rationality—people make judgment errors and depart from expected utility theory. Existing work in behavioral law and economics has examined the implications for employment law of the second and third bounds, as detailed in the sections "Bounded Self-Interest and Minimum Wage Regulation" and "Bounded Rationality, Employment Discrimination Law, and Employment Mandates" below. By contrast, bounded willpower has received very little attention in behavioral economics analysis of employment law, so I will begin there.

As the section "Bounded Willpower, Wage Payment Law, Pension and Social Security Regulation, and Age Discrimination Law" below describes, a surprisingly diverse set of employment law rules may be illuminated by considering bounded willpower. Bounded willpower suggests that individuals often greatly—"quasi-hyperbolically"—discount the future, and an important and much-studied implication of such behavior is that at any given point in time, individuals will fail to put adequate funds aside for their retirement even though their preferred *plan* would entail such saving. How does employment law respond to this disjunction between plans and actions?

One response it gives is to make some retirement saving from employees' earnings a mandatory feature of employment, as occurs through the Social Security system (which is discussed more fully below). But other employment law mechanisms are more subtle—and more directed to encouraging, rather than compelling, retirement saving by individuals with bounded willpower. The section "Wage Payment Law" suggests that this form of law supports the reliability of compensation in part through bonus payments, which are far more likely than ordinary wages to generate substantial retirement saving by individuals with bounded willpower. In addition, the regulation of employer-provided pension plans under the Employee Retirement Income Security Act (ERISA) targets bounded willpower both directly—through restrictions on early withdrawals from pension plans—and indirectly—through incentives provided to corporate executives, as described in the section "Pension Regulation." Finally, the section "Age Discrimination Law" provides an account of how age discrimination law may encourage retirement saving among individuals with bounded willpower by facilitating the use of back-loaded wage profiles, which place limits on the level of liquidity-constrained employees' present consumption.

The section "Bounded Self-Interest and Minimum Wage Regulation" describes evidence that employers and employees are frequently engaged in a "fairness dynamic" as a result of bounded self-interest. In the

fairness dynamic, employers choose to pay employees more than the minimum amount those employees would accept in exchange for their labor, and employees respond to such "fair behavior" by working harder than they otherwise would. The fairness dynamic turns out to have a number of implications for employment law's minimum wage regulation (Jolls, 2002). An interesting feature of the discussion in this section is that behavioral economics—although it is often viewed as comparatively more supportive than neoclassical economics of greater legal regulation—may at times carry a deregulatory impulse.

The section "Bounded Rationality, Employment Discrimination Law, and Employment Mandates" describes how both judgment errors and departures from expected utility theory have been brought to bear on the analysis of important features of employment law. "Erroneous" judgments (a concept that will be developed below) are relevant to understanding both the effects of existing employment discrimination law (Jolls, 2007a) and the desirability of proposed reforms of that law (Krieger and Fiske, 2006; Fiske and Krieger, this volume). Meanwhile, with respect to departures from expected utility theory, behavioral economics analysis of the "endowment effect" (Thaler, 1980) highlights the importance of the presence versus the absence of particular legally mandated employee benefits—such as health insurance and workplace leave—to equilibrium outcomes; in the presence of the endowment effect, the fact that a particular employee benefit is not contracted for, even in a market with perfect information, does not lead to the neoclassical economic prediction that mandating the benefit operates effectively as a tax that will depress employment levels (Jolls, Sunstein, and Thaler, 1998).

As described in this essay, a great many important features of employment law—from pension regulation to minimum wages to prohibitions on employment discrimination to mandated health insurance and workplace leave—are illuminated by behavioral economics. In some cases, behavioral economics analysis seems to produce a better fit with existing legal rules than does familiar neoclassical economic analysis; employment law rules that tend to be viewed critically by neoclassical economic analysts appear more sensible when viewed through a behavioral lens. At other times, as in the work of Krieger and Fiske (2006; Fiske and Krieger, this volume), the behavioral economics perspective suggests existing legal shortcomings that only come into focus through the adoption of this perspective. Both in understanding and in improving employment law, then, behavioral economics has an important role to play.

Bounded Willpower, Wage Payment Law, Pension and Social Security Regulation, and Age Discrimination Law

A large-scale study by Merrill Lynch asked baby boomers, "What percentage of your annual household income are you now saving for retirement?" and "What percentage of your annual household income do you think you should save for retirement?" The average gap between the two answers was 11% of household income (Bernheim, 1995).[1] Why might people choose to spend their earnings despite their stated desire to save for retirement?

Behavioral economics analyzes the disjunction between intentions and actual behavior by emphasizing the concept of bounded willpower—people's inability to stick to plans they set for themselves. Much empirical evidence supports the idea of such bounded willpower, as discussed below. Perhaps unsurprisingly, then, the law makes a wide range of attempts to increase people's retirement saving in the face of their apparent bounded willpower. Some of the law's attempts, such as the facilitation of Individual Retirement Accounts (IRAs), occur wholly outside of the employer-employee relationship, but others occur through regulation of this relationship and, thus, are discussed below. Behavioral economics illuminates how a diverse set of employment law rules—some having no direct relationship to retirement saving—are likely to increase such saving among individuals with bounded willpower.

Before proceeding, it is important to distinguish the discussion here from the existing literature on actual and potential legal responses to boundedly *rational* behavior—especially in the form of status quo bias and the resulting influence of default options—in the retirement saving context. Benartzi and Thaler (2007) provided a comprehensive recent survey of how bounded rationality shapes retirement saving behavior. The focus here, by contrast, is on bounded willpower.

Bounded Willpower

The problem of bounded willpower arises in a wide range of domains. People may choose to consume desserts over salads, start new projects rather than finishing old ones, and fail to go to the gym regularly despite their earnestly laid plans to do so. The discussion of bounded willpower here, however, focuses on the specific domain of retirement saving.

A study by Richard Thaler (1981) provided an early suggestion of the strong impatience that many

people display for immediate over delayed financial rewards. Thaler asked subjects to imagine that they had won $15 in a lottery and could either take the money now or put it away for later. The subjects were asked how much they would require for waiting to be as attractive as immediate payment, with time horizons of 1 month, 1 year, and 10 years. Subjects were specifically instructed to assume that the money would be preserved in a risk-free setting with no chance of future nonpayment. The median amounts stated for the 1-month, 1-year, and 10-year periods were $20, $50, and $100 respectively. These answers imply average annual discount rates of 345%, 120%, and 19% for the 1-month, 1-year, and 10-year periods.

Frederick, Loewenstein, and O'Donoghue (2002) provided a graphical illustration of the evidence from a range of empirical studies of discount rates that, as in Thaler's study, declined with the time horizon. The vertical axis on their graph shows the discount factor, and the horizontal axis shows the time horizon. The graph shows that the longer the time horizon, the higher the discount factor—and thus the lower the discount rate. Frederick, Loewenstein, and O'Donoghue further showed that the pattern of declining discount rates with the length of the time horizon is almost solely a product of people's strong impatience for near-term rewards; when they omitted studies with time horizons of less than one year from their analysis, discount rates and the length of the time horizon across the remaining studies were essentially uncorrelated.

The evidence discussed by Frederick, Loewenstein, and O'Donoghue points strongly toward a pattern under which periods after the present are discounted substantially in relation to the present but are discounted only modestly in relation to other periods. Mathematically, an individual's discount factor—the weight attached to utility in period k—may be approximated by $D(k) = \beta\delta^k$ (for $k > 0$ and $\beta, \delta \in (0,1)$), where β reflects the discounting of all periods other than the present and δ reflects the successively higher discounting of periods further into the future. This form of discounting is called quasi-hyperbolic.

Further empirical support for the sort of discounting discussed here comes from the observation of preference reversals in intertemporal decision making. A preference reversal occurs when an individual prefers to receive (say) $110 a week after a specified future date to $100 on this date but then, when the date actually arrives, prefers to receive $100 immediately to $110 in a week. Such inconsistency over time is an obvious consequence of the asymmetric discounting of future periods depending on whether they are being compared to other future periods or to the present. Frederick, Loewenstein, and O'Donoghue (2002) referred to a number of empirical studies that revealed such preference reversals.

With hyperbolic discounting, an individual will tend to defer saving in favor of present consumption in each period, even though such deferral is inconsistent with what the individual would have wanted to do in that period if the individual had been able to make an earlier choice about behavior in that period. Indeed, in some cases it is possible to show (Phelps and Pollak, 1968, p. 196 n1) that a pattern in which the individual is unconstrained in the consumption-saving decision in each period is *Pareto inferior to*—that is, worse for the individual in every period than—a pattern in which the individual is constrained to save in every period. Intuitively, all of the individual's temporal selves can be made better off through a commitment of each self not to give in to present desires to consume.[2]

Many responses to bounded willpower in the retirement saving context are possible, but one intriguing potential response that operates without any form of direct legal constraint—a theme to which I will return below—is the recharacterization of employees' earnings through various forms of mental accounting. Some earnings do not appear to be coded as "available for present consumption" in the same way that ordinary wages are. People appear far more likely, for example, to save substantial fractions of bonus payments, even when they fail to engage in substantial saving out of funds received as ordinary wages (Thaler, 1990). Contrary to the assumption of neoclassical economics, not all dollars are treated the same (Zelizer, 1994). Thaler and Shefrin (1981) referenced the example of Japan, where bonus schemes are common and retirement saving rates are high. Links between this phenomenon and employment law will be explored below.

Implications of Bounded Willpower for Wage Payment Law, Pension and Social Security Regulation, and Age Discrimination Law

This subsection discusses a diverse range of employment law rules in light of bounded willpower in the retirement saving context.

WAGE PAYMENT LAW

As just noted, many people save substantially more out of bonuses than out of regular wage payments (Thaler, 1990). However, an important issue with compensating employees in part through bonuses is that, relative to ordinary wages, bonuses can be highly unreliable; precisely the feature that makes bonuses

more conducive to retirement saving—their lumpy nature, by contrast with the payment of wages at more frequent intervals—makes them less likely actually to be paid because employees may have limited leverage against their employers by the time bonuses become due. Indeed, even with ordinary wages, employers historically would sometimes refuse to pay employees for work done, notwithstanding the fact that the frequency of ordinary wage payments meant employees might have the opportunity to stop working in response to the nonpayment of wages. With bonuses, by contrast, typically all of the work has been performed by the time the lump-sum bonus becomes due.

Employment law's response to the danger of withheld wages, bonuses, and other forms of compensation came in the form of wage payment statutes, which are common across the United States. These statutes require the regular payment of wages and provide employee-friendly procedures and remedies to ensure that the right to payment of earned wages is protected. Whereas in the absence of wage payment statutes an employee's remedy for unpaid wages would lie purely in contract—with the recovery equal merely to the owed wages and the employee responsible for paying an attorney to bring suit—wage payment law creates a penalty system designed to give employers adequate incentives to make regular wage payments. Employers who violate the statutes may be liable for liquidated damages or even criminal sanctions as well as being responsible for the cost of the employee's attorney (e.g., Wash. Rev. Code ch. 49.48). In many states, wage payment law embraces bonuses (at least under some circumstances) as well as ordinary wages (e.g., *Gurnik v. Lee*, Ind. Ct. App. 1992); in this way, employment law supports retirement saving by individuals with bounded willpower.

Note that the discussion here is descriptive rather than normative; it concerns the effects of wage payment law on retirement saving, not the normative desirability of such law. From a normative perspective, a general concern with legal support for compensation through bonuses as a mechanism for encouraging retirement saving is that, absent the possibility of a Pareto improvement for all of an individual's multiple selves (a prospect briefly noted above), it is unclear whether increasing retirement saving is, on balance, normatively desirable; assuming, plausibly, that such an increase often benefits future selves at the expense of earlier selves, an argument in favor of that ordering is needed.

Beyond this general issue, a potential concern with legally supporting bonus-based compensation as a way of encouraging retirement saving is that, precisely because individuals account for bonuses in a way different from the way they account for ordinary wages, it is possible that employers are more able to cut, or fail to increase, compensation than they would be able to do if the compensation were paid solely in the form of ordinary wages. Kahneman, Knetsch, and Thaler (1986) presented the following two questions to survey respondents:

> Question 6A. A small company employs several people. The workers' incomes have been about average for the community. In recent months, business for the company has not increased as it had before. The owners reduce workers' wages by 10% for the next year.
>
> Question 6B. A small company employs several people. The workers have been receiving a 10% annual bonus each year and their total incomes have been about average for the community. In recent months, business for the company has not increased as it had before. The owners eliminate the workers' bonus for the year.

While only 39% of respondents presented with the first question found the wage reduction acceptable, 80% of respondents presented with the second question found the elimination of the bonus acceptable. In the present context of retirement saving, because a substantial fraction of bonus payments goes to saving—which by hypothesis is undervalued by individuals' present selves—employers relying on bonuses may over time face lower total compensation demands from their employees than employers relying solely on ordinary wages.

PENSION REGULATION

An obvious mechanism for retirement saving is employer-sponsored pension plans. The central employment law governing such plans is ERISA, which imposes various requirements on these plans in order for the plans to qualify for favorable federal income tax treatment. Because of the magnitude of the tax advantages, in practice employer-sponsored pension plans tend to conform to ERISA's requirements. The discussion here will focus on 401(k) plans, which are individual employee retirement accounts and which represent the major vehicle through which most employees today participate in employer-sponsored pension plans.

The most direct prediction of bounded willpower is that individuals will often be tempted to withdraw 401(k) funds for current consumption. To be sure, such withdrawals might be desirable in some cases even for an individual with unbounded willpower, as an emergency might have arisen. But an individual

with bounded willpower will be tempted to make withdrawals even apart from such exigencies.

ERISA's restrictions on employees' ability to make early withdrawals from 401(k) plans can be understood as a natural response to bounded willpower. (Weiss, 1991, offers related discussion.) Under ERISA, employees who wish to make early withdrawals typically must establish that they fall into a set of categories structured around either exigent circumstances or alternative forms of earning or saving (see Treas. Reg. §1.401(k)-1(d)). The first category includes preventing eviction or home foreclosure and covering major medical expenses; the second category includes pursuing higher education (which generally increases earnings) and purchasing a primary home (an alternative form of saving). Interestingly, individuals asked whether they would prefer to face fewer restrictions on withdrawals generally reported that they would not prefer fewer restrictions (Laibson et al., 1998). ERISA does allow 401(k) plans to permit borrowing by individuals against the plan proceeds,[3] although such borrowing is probably made less likely by the fact that in mental accounting terms the funds have been "marked" for long-term purposes (Thaler, 1994). Finally, ERISA is a mandatory regime and, thus, cannot be avoided by a contrary agreement between employer and employee; otherwise, much of the benefit of ERISA in responding to the problem of bounded willpower could disappear because profitable renegotiation of limitations on withdrawals would exist at times at which individuals are tempted to spend (although the mental accounting point might be enough to prevent such an outcome). At bottom, then, ERISA sets up a flexible regime—more flexible, for instance, than the sort of uniform tax discussed by Beshears et al. (2005) or than illiquid assets that are costly to unload in exigent circumstances—that nonetheless helps to support retirement saving by individuals suffering from bounded willpower.[4]

An additional aspect of ERISA that responds to the problem of insufficient retirement saving under bounded willpower is the way in which the statute harnesses the personal incentives of corporate executives in the service of increased retirement saving. Because of the high income and wealth of many corporate executives, putting aside funds for retirement will typically be less difficult for these individuals than for individuals at the low end of the wage scale. (Note that the point is not that corporate executives have fewer bounds on their willpower as a general matter. Plenty of corporate executives have as much trouble sticking to their exercise plans as the rest of society does. The point here is simply that it will typically take less willpower to put aside money for retirement when one is earning at a high level than when funds are scarce.) Thus, a reasonable worry with respect to retirement saving is that those at the helm of firms will not have sufficient incentives to structure 401(k) plans in ways that provide the broadest possible support for saving by individuals with bounded willpower. In response to this worry, ERISA limits retirement contributions by high-level employees unless low-level employees are participating to an adequate degree (see, e.g., Bankman, 1988). These limitations encourage corporate executives to think creatively about structuring 401(k) plans to encourage saving in a robust way (Thaler, 1994).

In a clear example of the effects of this aspect of ERISA, Thaler and Benartzi (2004) described how executives at one company sought to increase low-level employees' saving (in order to increase the executives' own retirement saving options) through a 401(k) plan structure called Save More Tomorrow. Under this plan, which was developed by Thaler and Benartzi, individuals are invited to save a fraction of future pay (often, but not always, taken from future pay raises). Because individuals are not being asked to reduce their current consumption in any way, bounded willpower is less likely to interfere with a decision to save. In fact, the Save More Tomorrow plan has produced striking increases in retirement saving at some early-adopting companies (Thaler and Benartzi, 2004; see also Benartzi, Peleg, and Thaler, this volume). At one company, for example, those who participated in the plan—the vast majority of employees—more than tripled their saving rates in 28 months.

As with respect to the effects on retirement saving of covering bonuses under wage payment law, the normative analysis here is complex. Once again it is possible that the law benefits future selves at the expense of earlier selves, with the attendant normative question of the desirability of that outcome. Alternatively, it is possible that—precisely because individuals' present selves show limited concern with retirement saving—individuals will care less about their compensation levels with a Save More Tomorrow or similar plan in place. In other words, if such a plan were in place, an individual might have a less negative reaction to getting a 3% raise instead of a 5% raise because much of the pay increase would be going to retirement saving instead of present consumption. Regardless of the normative analysis, however, what is clear is the descriptive point that ERISA's nondiscrimination rules, in encouraging steps such as the adoption of Save More Tomorrow plan, are likely to have the effect of increasing retirement saving.

SOCIAL SECURITY

Both wage payment law and ERISA impose various mandatory terms in employer-employee relationships—terms that I have suggested tend to have the

effect of increasing retirement saving by individuals with bounded willpower. But these mandatory terms operate in conjunction with voluntary choices made in the employer-employee relationship; employers need not use employee bonuses or offer pension plans at all, but if they do, then mandatory rules under wage payment law and ERISA attach.

A different type of employment law mandate that also may be understood as a response to bounded willpower is the Social Security system. Social Security is, in magnitude terms, a very significant aspect of the "employment contract" between employers and employees; those who work—and only those who work—make contributions to the program, retirement benefits are limited to the employees who contributed and to the dependents of these employees, and the dollar amounts involved are very large (Feldstein and Liebman, 2002).

Employer-employee agreement is always insufficient to avoid the requirements of the Social Security system; in any employment relationship, Social Security payroll deductions must be made. And, in contrast to the case of employer-provided pension plans under ERISA, with Social Security there is no opportunity for early withdrawal or borrowing, no matter how exigent the circumstances. Thus, Social Security can be (and commonly is, see, e.g., Weiss, 1991; Feldstein and Liebman, 2002) understood as an employment law response to bounded willpower in the retirement-saving context.

Once again, it is important to distinguish the descriptive claim that employment law (here in the form of Social Security) facilitates retirement saving by individuals with bounded willpower from the normative claim that this form of regulation is desirable. Even if one is willing to average gains and losses of different temporal selves (abandoning a Pareto standard), adopting a Social Security system in response to bounded willpower may or may not increase societal welfare (Amador, Werning, and Angeletos, 2006; Imrohoroglu, Imrohoroglu, and Joines, 2003; Kumru and Thanopaulos, 2008). The descriptive claim that Social Security increases retirement saving in a world of bounded willpower, while not completely uncontroverted, is on relatively firmer ground.

AGE DISCRIMINATION LAW

An additional potential means of facilitating retirement saving in a world of bounded willpower is the back-loading of wages. If funds do not arrive until later in the life cycle, individuals are effectively forced not to spend them earlier (except in the event that they are able to borrow against the back-loaded amounts—something we observe only to a limited degree in practice).

In fact, substantial empirical evidence suggests that a pattern of back-loaded wages—wages that slope upward with age even after controlling for changes in productivity—exists for many employees (e.g., Medoff and Abraham, 1980, 1981). Explanations for the apparent appeal of back-loaded wages include not only the bounded willpower emphasized here but also incentive-based explanations (Lazear, 1979) and psychological explanations rooted in individuals' desire to experience gains over time (Frank and Hutchens, 1993; Loewenstein and Sicherman, 1991). Note that the linking of back-loaded wages to bounded willpower requires that individuals cannot renegotiate their wage levels during periods of temptation to spend—an assumption that may be reasonable in many cases because a wage change would not be reflected until the individual's next paycheck at the earliest. (See Laibson et al., 1998, for a related discussion.)

An important problem with back-loaded wages, however, is that they are highly vulnerable to exploitation by employers in the absence of effective legal constraints; at the end of the life cycle, when the back-loaded portion of employees' compensation comes due, employees will be net drains on employers, and therefore employers will be eager to discharge them if possible. Such cost-based discharges are in fact a frequent feature of litigated age discrimination cases. (See Jolls, 1996, for examples and further discussion.) Thus, the legal limitations on the discharge of older workers under the Age Discrimination in Employment Act (ADEA) may facilitate reliance on back-loaded wages. (See Neumark and Stock, 1999, for corroborating empirical evidence.)

It bears noting, however, that if back-loaded wages are desired *solely* because of bounded willpower and not in part because of the incentive or psychological considerations noted above, then alternate forms of legal intervention might be the optimal response. In particular, back-loaded wage entitlements could be packaged as vested portable pensions, so that older employees would not present higher current wage costs for employers (Jolls, 1996). With the removal of such higher wage costs, the employer opportunism problem noted above would disappear. In the absence of vested portable pensions, however, support for back-loaded wage structures through the ADEA will tend to encourage retirement saving among individuals with bounded willpower.

Once again, the analysis here is descriptive in nature. A normative account of the ADEA as a response to bounded willpower in the retirement saving context would be subject to the now-familiar caveats about the conflict between maximizing the satisfaction of earlier versus later selves' preferences and about the potential effects on employees' overall level of compensation.

Bounded Self-Interest and Minimum Wage Regulation

Turning from bounded willpower to bounded self-interest, behavioral economics has emphasized the way in which individuals often choose actions that do not maximize their material self-interest. Substantial empirical evidence suggests that in the employment context, employees often respond to "fair" wages by working harder than they otherwise would (even absent any mechanism for adjusting wages in response to the effort level). The present section first describes this "fairness dynamic" and then discusses its implications for minimum wage regulation.[5]

The Fairness Dynamic in Employer-Employee Relationships

As Robert Solow has written (1990, pp. 9–10), "The fundamental reason for believing that fairness is a factor in labor markets is what we know about our own society and culture. . . . Wage rates and employment are profoundly entwined with social status and self-esteem." Indeed, fairness may play many important roles in wage setting and other aspects of the employment relationship. Studies by Kahneman, Knetsch, and Thaler (1986), Blinder and Choi (1990), and Campbell and Kamlani (1997), for example, examined perceptions of the fairness or unfairness of wage adjustments in response to various demand- or supply-side shifts in the economy and found that such perceptions had significant effects. Fairness also appears to play a major role in the determination of the relative wages of various groups of employees within a firm, as Levine (1993) and Bewley (1999, pp. 75–82), among others, have emphasized.

The discussion below, however, focuses not on this whole range of fairness behavior in the employment relationship but rather on one specific form of such behavior. That behavior has its theoretical basis in the efficiency wage model of Akerlof and Yellen (1990). In this model, employers pay wages above employees' "reservation wage"—the minimum level they would demand for their services—in order to induce reciprocation in the form of high levels of effort. On a macroeconomic level, this fairness dynamic can explain the otherwise puzzling existence of involuntary unemployment in the economy.[6]

Almost a century ago, Sumner Slichter (1929) emphasized the role of fair treatment in spurring high levels of effort by employees. Slichter noted that the poor state of the economy beginning in 1920 had not led to a reversion to the harsh labor practices that prevailed in the "buyers' market" for labor before the First World War, and he concluded that "possibly the most important determinant of post-war labor policies . . . has been the growing realization by managers of the close relationship between industrial morale and efficiency" (pp. 396–397, 401).

Fehr, Kirchsteiger, and Riedl (1993) provided strong empirical support for the Akerlof and Yellen version of the efficiency wage model. The authors' experimental results have been replicated in numerous subsequent studies, including one in which the stakes were two to three times the participants' monthly incomes.[7]

In the first stage of Fehr, Kirchsteiger, and Riedl's experiment, "employers" were given a specified period of time in which to bid for the services of a single, unknown "employee."[8] Bids consisted of the wage that the employer would pay the employee. In the second stage, those employees who had accepted offers of employment at the specified wages were able to set an effort level at which they would perform. Higher effort levels were associated with increases in employers' payoffs, because employers earned higher profits, but with decreases in employees' payoffs, because effort was costly. Wages were not made contingent upon effort levels, and employers had no ability to retaliate for low effort levels in future periods because they did not know the identity of their particular employee. Thus, it was impossible for employers to induce high effort levels by a strategy of monitoring employees and punishing them for poor performance.

According to the traditional model, the results of this experiment would be quite predictable. Employees will always choose the minimum effort level in the second period so as to maximize their payoffs; their wage has been fixed in the first period, punishment for low effort is not feasible, and effort is costly. Employers, aware of this incentive, should assume low employee effort and offer a wage that puts employees just above their "reservation level" (the minimum level they would demand for their services). Employees should accept the offered wage since it is above the reservation level. The result is a low-wage, low-effort equilibrium. Does this simple prediction square with the experimental results? No. Employers in the above setting typically chose wage levels above the level predicted by the analysis just described, and employees responded by choosing effort levels significantly in excess of the minimum feasible level.

These results suggest concerns with fairness. Workers who receive wages above the low level predicted by the traditional analysis may offer high levels of effort in response based on their perceptions of the fairness of the employers' behavior, and employers, aware of this result, can maximize their profits by offering such generous wages. This is the basic mechanism

contemplated by the Akerlof and Yellen theory. Subsequent work by Fehr et al. (1998) confirmed the fit between the Akerlof and Yellen model and the behavior observed in the experiments by showing that employers' offers of high wages do not reflect an unwillingness by employees to work for less but instead, as envisioned by the efficiency wage model, reflect a desire by employers to encourage high levels of effort by paying employees more than the reservation level they would demand for their services.

Considerations of fairness arise in the context under discussion because when employers cannot directly monitor their employees' effort (as in the experiments here), employers seek to encourage employees to perform well in response to being offered "fair" wages. Where high effort cannot be ensured through monitoring and punishment, a fair wage provides an alternative means by which an employer may be able to encourage an employee to exert effort.

In broad terms, the fairness dynamic described here is consistent with the literature in economics and political science suggesting the efficiency aspects of "trust" relationships. Empirically, there is some evidence that higher levels of trust are correlated with better economic performance across regions and across countries (e.g., La Porta et al., 1997). These results suggest that the opportunity to build upon trust relationships enhances efficiency.

Implications of the Fairness Dynamic for Minimum Wage Regulation

At the most basic level, the fairness dynamic suggests that a minimum wage requirement may be less necessary to raise wages than might otherwise be thought, as the essential idea behind the dynamic is that employers and employees may find their way to an equilibrium with higher wages entirely on their own. But at some level this observation is too simple, for one premise of the fairness dynamic is that high effort cannot be ensured by the direct mechanism of monitoring effort and then punishing employees who fail to perform up to par. Such monitoring and punishment are obviously possible in some settings, and thus a more refined set of conclusions from the fairness dynamic focuses on settings in which a minimum wage requirement is likely to be more or less necessary to raise wages.

The discussion to follow emphasizes the ease of monitoring rather than the ease of punishment for low effort by an employee because the former seems easier to theorize about a priori.[9] The discussion uses differences in the likely ease of monitoring to try to make sense of the scope of coverage of the minimum wage requirement of the Fair Labor Standards Act (FLSA) and to predict variations in the degree of compliance with this requirement within covered sectors. The basic insight is that once fairness is taken into account, a minimum wage requirement is less necessary to raise wages, all else equal, in situations in which employees are difficult to monitor than in situations in which they are relatively easy to monitor. (A minimum wage may still be important in setting expectations of what counts as a "fair" wage, however.) If employees are difficult to monitor, then fairness considerations may push toward a higher wage wholly apart from legal regulation, as employers strive to pay employees "fairly" in order to encourage diligence and hard work on the employees' part. If, by contrast, monitoring is relatively easy, then fairness considerations do not create any upward pressure on wages because employees can simply be fired if monitoring discloses that they have not performed well. Minimum wage laws are more necessary to raise wages, all else equal, in the latter context.

In terms of the FLSA's coverage, the claim here will not be that the fairness dynamic provides a comprehensive framework to make sense of the overall statutory structure of the FLSA's minimum wage requirement. That requirement is subject to a number of rather random-sounding exemptions, including for various employees working in the fishing and agricultural industries, employees working in summer camps and similar recreational establishments, and employees employed by small newspapers or telephone companies.[10] The analysis offered here does not purport to explain all of these exemptions, just to make some sense of the specific ones discussed below.

THE HISTORICAL EXEMPTION OF DOMESTIC SERVICE EMPLOYEES

Until 1974, all domestic service employees were exempt from the FLSA (Smith, 2000). At one level, this exemption seems quite surprising, as at least some domestic employees are quite vulnerable as economic actors. Why should these employees have been excluded from the coverage of the minimum wage requirement?

Concerns of family privacy have been adduced in support of the exemption of domestic service employees (Smith, 1999). (Other accounts emphasize racial aspects, as noted below.) As one historical source put it, "[The domestic's] position is peculiar. She is in the family, but not of it."[11]

The fairness dynamic, however, provides an interesting variation on this theme of household "privacy." Some forms of household work—for instance, care for children—are difficult to monitor. While one knows whether the employee is present for work, the

quality of the work is, or can be, extremely subtle in its variations, in ways that cannot be monitored well unless the employer hovers over the employee, which of course would tend to defeat the purpose of hiring the employee in the first place. The fairness dynamic suggests that employers and employees may end up at an equilibrium with a higher-than-expected wage, and a correspondingly higher level of effort, without the intervention of a minimum wage requirement. If this analysis carries some truth, then a minimum wage requirement may be less necessary to raise the wages of certain domestic service employees than to raise the wages of otherwise similar employees working in different settings.

Of course, many domestic service employees perform tasks—such as various housework duties—that may not involve the sort of discretion associated with child care, and much of the literature on domestic service employees and their abuse at their employers' hands focuses directly on such employees, who are not the subject of the fairness argument here and who may very well desperately need the protection of a minimum wage requirement (e.g., Smith, 2000). Moreover, at the other end of the spectrum, certain domestic service employees—such as high-level professional nannies—are in a different category from those domestic service employees who could conceivably stand to gain from the application of a minimum wage requirement, as high-level professional nannies earn dramatically in excess of the minimum wage (see, e.g., Eaton, 1998). However, as described by Jolls (2002), some in-home child-care workers do earn relatively low wages (and presumably also did in the past, although it is hard to get access to good data for the pre-1974 period for child-care workers as distinguished from other domestic service employees); thus it remains an interesting question whether it makes sense for the minimum wage requirement to apply to these child-care workers.

It is important to emphasize that the notion of a "higher wage" equilibrium as a result of the fairness dynamic does not necessarily ensure that the employees in question were earning—prior to the elimination in 1974 of the FLSA exemption—a "living wage," that is, one capable of sustaining them at reasonable standards. Even a wage above the minimum required by the FLSA might well not be a living wage; whether it is depends, of course, on the gap between the legally required minimum and the level required for a living wage. As an interesting point of comparison, in the Fehr, Kirchsteiger, and Riedl study (1993) described above, the result of fairness behavior is an average wage that is more than twice that predicted by the traditional economic theory.

Note that the point here is not that Congress drafted the exemption for domestic service employees based on the fairness dynamic described above. My point here is not to describe the intent or goals of Congress. Instead, the fairness dynamic provides a possible rationalization, or way to make sense, of the statutory exemption of domestic service employees, whose exclusion Linder (1987, 1992, pp. 154–155) suggested in fact resulted from racism on the part of New Deal lawmakers.

THE FAILURE TO COVER INDEPENDENT CONTRACTORS

The FLSA's minimum wage requirement applies to "employees" but not to "independent contractors." Unlike the limit pertaining to domestic service employees, this limit on the coverage of the FLSA continues in effect today. As with the aspects of the FLSA discussed above, it may be possible to make some sense of this feature of the law by reference to the fairness dynamic and the relative difficulty of monitoring independent contractors versus employees.

Under the FLSA, whether an individual is an independent contractor or an employee turns on the following factors:

1. The nature and degree of the employer's control as to the manner in which the work is to be performed
2. The individual's opportunity for profit or loss depending upon his managerial skill
3. The individual's investment in equipment or materials required for his task, or his employment of workers
4. Whether the service rendered requires a special skill
5. The degree of permanency and duration of the working relationship
6. The extent to which the service rendered is an integral part of the employer's business[12]

The first, fourth, and fifth of these factors are likely to correlate with the difficulty of monitoring an individual's work. The less control an employer has as to the manner in which the work is to be performed (the first factor), the more difficult it may be for the employer to monitor that work. Similarly, the more skilled the individual's work (the fourth factor), the more difficult it may be for the employer to monitor the work. And finally, the lesser the degree of permanency and duration of the working relationship (the fifth factor), the greater the difficulty of (successful) monitoring of the individual's work, as there will not be a long horizon over which the employer can look for poor performance. Based upon these factors, the

work of independent contractors may be more difficult, all else equal, to monitor than that of employees, and thus, according to the fairness dynamic, the application of a minimum wage requirement will be less necessary, all else equal, to raise the wages of independent contractors than to raise the wages of employees.

The fairness dynamic thus provides some assistance in making sense of the oft-criticized failure of the FLSA to cover independent contractors. This is not to say, though, that every exclusion accomplished by that coverage failure makes sense; some exclusions, such as that by some courts of migrant farm workers, seem hard to consider sensible or warranted.

COSTS OF MINIMUM WAGE LAW

The central implication of the fairness dynamic is that the minimum wage requirement of the FLSA is less necessary, all else equal, to raise wages in settings in which monitoring is difficult than in settings in which monitoring is less difficult. But perhaps this argument implies nothing more than that a minimum wage requirement would simply be irrelevant in settings in which, because of monitoring difficulties, fairness pushes up wages without the need for legal intervention. What are the costs, if any, of imposing a minimum wage requirement? Why bother exempting certain employees if the law would simply be irrelevant to them given the operation of the fairness dynamic?

From a law and economics perspective, it may seem obvious that any form of legal regulation is likely to carry with it costs, so that a regulation that is believed to produce no or few positive effects obviously should not be imposed. But it is worth pausing briefly to consider what exactly these costs might be insofar as minimum wage regulation is concerned.

First, like any legal regulation, a minimum wage requirement imposes administrative costs, for even an employer who has conformed substantively to the requirement may always be haled into court and asked to prove to the court's satisfaction that it has done so. The associated legal and other costs may be substantial. Furthermore, an employer who must be able to prove in court that it has met the minimum wage requirement will need to track and maintain records of the specific number of hours worked by each employee in exchange for the pay received by the employee, and this practice obviously entails costs. Most related to the ideas explored above, it may be the case that minimum wage regulation in a particular setting would serve as a signal to market participants that employers are not sufficiently trustworthy to be left on their own in setting wages. Minimum wage regulation thus might disrupt the operation of the fairness dynamic.

POLITICAL ORIENTATION

As suggested at the beginning of this essay, it is interesting to observe that the policy implications of the fairness dynamic tend to be distinctly of the laissez-faire variety. If people will behave appropriately without legal regulation—as the fairness dynamic suggests they may—then perhaps the market should be left to function without legal regulation. This creates an intriguing political juxtaposition, as political liberals are probably more open in general to the importance of a phenomenon like fairness, but then when one looks to implications for the law it turns out that, at least in this context, the conclusions are generally more apt to please political conservatives.

While some might naturally assume that behavioral economics (as compared to traditional economic theory) is more, rather than less, likely to provide normative support for legal intervention—and while in some cases, such as in the discussion of judgment errors and employment discrimination law, below, this may be true—the case of fairness is an important counterexample. If we take seriously the idea that people care about fair treatment, they may be more likely than we would otherwise assume to resolve their conflicts on their own, and the role of the law will accordingly be reduced.

Two qualifications to this statement are important. First, an implicit assumption underlying the laissez-faire nature of the normative conclusion just outlined is that the benefit of pushing up wages outweighs the cost of the reduced employment that is likely to come along with higher wages for those who remain employed. When a minimum wage is imposed by Congress, one might reasonably assume that the trade-off between higher wages and higher employment has been resolved by the polity in favor of higher wages (assuming that there is in fact such a trade-off). But when the increase in wages occurs, as in the discussion here, through the operation of market forces rather than through legislation, it is, ironically, possible at least in theory that the resulting wage is too high relative to the social optimum, and thus that government intervention is needed to protect opportunities for employment from encroachment by excessive wage levels. So fairness, in this particular context, could conceivably argue for the necessity of market intervention rather than against the necessity of such intervention.

Second, it is possible that the occurrence of the fairness dynamic may turn on cultural or other similarities within the workplace. Trust may not be able to cross cultural barriers, and if so, legal intervention may remain necessary to achieve desirable outcomes.

The more general point is that it is often more difficult than observers have realized to generalize about the political orientation of behavioral law and economics.

Bounded Rationality, Employment Discrimination Law, and Employment Mandates

This section addresses employment law responses to a third bound on human behavior, bounded rationality. Because of the breadth of the category of bounded rationality, it is useful to subdivide this category into the subcategories of judgment errors and departures from expected utility theory.

Judgment Errors and Employment Discrimination Law

This subsection begins by describing the nature of judgment errors and then discusses the ways in which concerns about judgment errors both are and are not well addressed by existing employment discrimination law.[13]

JUDGMENT ERRORS

One important general way in which human rationality is bounded is that people rely on mental shortcuts or rules of thumb—known as heuristics—that function well in many settings but lead to systematic errors in others. Consider, for instance, the well-known study involving people's judgments about a 31-year-old woman, Linda, who was concerned with issues of social justice and discrimination in college.[14] People tended to say that Linda was more likely to be a "feminist bank teller" than to be a "bank teller." This judgment is patently illogical, for a superset cannot be smaller than a set within it. The source of the mistake is the representativeness heuristic, by which events are seen to be more likely if they "look like" certain causes. In the case of Linda, the use of the representativeness heuristic leads to a mistake of elementary logic—the conclusion that characteristics X and Y are more likely to be present than characteristic X.

Research in cognitive psychology emphasizes that heuristics of this kind frequently work through a process of "attribute substitution," in which people answer a hard question by substituting an easier one (Kahneman and Frederick, 2002). For instance, people might resolve a question of probability not by investigating statistics, but by asking whether a relevant incident comes easily to mind. Often (although not always) the use of the heuristic occurs without any conscious awareness on the part of the actor; within the domain of "dual process" approaches (see, generally, Chaiken and Trope, 1999), heuristic-based thinking is typically rooted in System 1, which is rapid, intuitive, and error-prone, rather than in the more deliberative System 2.

An important category of System-1 thinking, which may be heuristic-based in an important sense, is implicit bias on the basis of race and other group-based traits. Such implicit bias is most familiarly associated with scores on the Implicit Association Test (IAT), which has been taken by large and diverse populations on the Internet and elsewhere (Greenwald, McGhee, and Schwartz, 1998; Nosek, Banaji, and Greenwald, 2002; see also Hardin and Banaji, this volume). The IAT asks individuals to perform the seemingly straightforward task of categorizing a series of words or pictures into groups. Two of the groups are racial or other categories, such as "black" and "white," and two of the groups are the categories "pleasant" and "unpleasant." In the version of the IAT designed to test for implicit racial bias, respondents are asked to press one key on the computer for either "black" or "unpleasant" words or pictures and a different key for either "white" or "pleasant" words or pictures (a stereotype-consistent pairing); in a separate round of the test, respondents are asked to press one key on the computer for either "black" or "pleasant" words or pictures and a different key for either "white" or "unpleasant" words or pictures (a stereotype-inconsistent pairing). Implicit bias against African Americans is defined as faster responses when the "black" and "unpleasant" categories are paired than when the "black" and "pleasant" categories are paired. The IAT is rooted in the very simple hypothesis that people will find it easier to associate pleasant words with white faces and names than with African American faces and names—and that the same pattern will be found for other traditionally disadvantaged groups. In fact, implicit bias as measured by the IAT has proven to be extremely widespread; most people tend to prefer white to African American, young to old, and heterosexual to gay (Greenwald, McGhee, and Schwartz, 1998; Nosek, Banaji, and Greenwald, 2002).

Implicit bias is System 1 in nature because it is largely automatic; the characteristic in question (skin color, age, sexual orientation) operates so quickly in the relevant tests that people have no time to deliberate. It is for this reason that people are often surprised to find that they show implicit bias. Indeed, many people say in good faith that they are fully committed to an antidiscrimination principle with respect to the very trait against which they show a bias (Greenwald, McGhee, and Schwartz, 1998). Not surprisingly, this sort of bias has complex implications for employment law rules regulating discrimination on the basis of race and other group-based traits, as discussed next.

IMPLICIT BIAS AND EMPLOYMENT DISCRIMINATION LAW

In some ways, implicit bias raises the possibility that existing employment discrimination law has bias-reducing effects beyond those usually contemplated; in other ways, implicit bias suggests important shortcomings of existing employment discrimination law.

The Bias-Reducing Effects of Existing Employment Discrimination Law

As summarized by Jolls (2007a), a substantial literature in employment discrimination law argues that existing law is severely misguided as a result of its failure to target implicitly biased behavior for legal prohibition. The central target of existing employment discrimination law is consciously, rather than implicitly, biased behavior. Nonetheless, Jolls (2007a) emphasized that existing law does have important effects on implicit bias because in prohibiting consciously biased employment actions, as well as in restricting harassing behavior in the workplace, existing law helps to reduce implicit bias in employment relationships.

Consider, first, the prohibition by existing employment discrimination law of consciously biased hiring decisions, firings, and other employment actions. This prohibition naturally tends to increase workplace diversity, and substantial evidence suggests that more diverse environments encourage lower implicit bias (Dasgupta and Asgari, 2004; Lowery, Hardin, and Sinclair, 2001; Richeson and Ambady, 2003). Thus existing employment discrimination law will tend to reduce implicit bias (Jolls, 2007a). Lowery, Hardin, and Sinclair (2001), for example, found that an in-person IAT administered by an African American rather than a white experimenter yields significantly lower measured levels of implicit bias. In other words, people's speed in characterizing black-unpleasant and white-pleasant pairs is closer to their speed in characterizing black-pleasant and white-unpleasant pairs when an African American experimenter is present. Similarly, Richeson and Ambady (2003) found that white test subjects paired with an African American partner exhibited less implicit bias as measured by the IAT than white test subjects paired with a white partner. These findings suggest that simply by increasing the level of population diversity in the workplace, existing employment discrimination law tends to reduce the level of implicit bias.[15]

A similar analysis applies to the prohibition by employment discrimination law of harassing behavior in the workplace. Both evidence and common sense suggest that the presence of stereotypic images of a particular group—for instance, a pin-up calendar featuring nude women in submissive poses—tends to increase implicit bias with respect to that group.[16] If so, then, by restricting negative or demeaning depictures of particular groups, existing harassment law helps to reduce the level of implicit bias (Jolls, 2007a).

The Limits of Existing Employment Discrimination Law

The suggestion above was that the asserted irrelevance of existing employment discrimination law to the phenomenon of implicit bias was overstated; existing law, although it does not aim at implicit bias in any direct way, nonetheless is likely to have the effect of reducing such bias in the workplace. Still, it is to be expected that the formulation of existing law without real attention to the problem of implicitly biased behavior would leave such behavior underregulated in important ways, and Krieger and Fiske's recent work developed a particularly significant respect in which this is true (Krieger and Fiske, 2006).

Krieger and Fiske observed that under existing law, an employer may defend against a claim of employment discrimination by establishing that when it made the challenged employment decision it was acting under an "honest belief" that the employee had a particular problem or flaw; the law is inattentive to the possibility that the perception of the problem or flaw may itself be the product of racial or other group-based bias. Thus, for instance, if an employer establishes in court that it terminated an employee for (what the employer perceived was) poor performance, the employer automatically prevails even though, as Krieger and Fiske noted, there is a real chance that the employer's perception was influenced by implicit bias (pp. 1036–1038). In one striking study they described, subjects needed to rank the importance of education and relevant job experience for choosing a high-level construction manager. In the study, when the male candidate had more education and less relevant job experience, subjects reported that they viewed education as more important than job experience, and most selected the male candidate. But when the male candidate had more job experience, subjects ranked job experience as more important, and again most selected the male candidate. The subjects' "honest belief" appeared to be that they were choosing based on either education or job experience, and indeed when subjects were required to rank the criteria before knowing how the male and female candidates fared, the gender bias largely disappeared. But in the real world, in which traits are known as information is being processed, Krieger and Fiske suggested that an "honest belief" may often bear a heavy racial or other imprint. Note that this argument is most applicable to cases of subjective "honest" reasons; objective reasons—for instance, the employee's attendance record—would be at least somewhat easier to pinpoint as biased if, indeed, they were.

As Krieger and Fiske noted, the scope of their argument is substantial because in a world in which antidiscrimination ideals hold strong sway, "people whose preferences are implicitly shaped by group membership spontaneously search for independent decision criteria consistent with their preference, and use those criteria to justify their choices to themselves and others" (p. 1037). Ironically, the very strength of the norms against discrimination in today's society encourages decision makers to think in ways that do not appear to them to be biased—even when they are.

Departures from Expected Utility Theory and Employment Mandates

A second type of boundedly rational behavior involves departures from expected utility theory. The following discussion first describes this category and then identifies how departures from expected utility theory help to predict the effects of a range of employment mandates, from health insurance mandates to family- and medical-leave mandates.[17]

DEPARTURES FROM EXPECTED UTILITY THEORY

Although expected utility theory is a foundational aspect of traditional economics, empirical departures from the precepts of this theory are common. A prominent example is the endowment effect, according to which individuals' behavior is influenced by their starting points. In the well-known mugs experiments (Kahneman, Knetsch, and Thaler, 1990), for instance, randomly selected individuals who received mugs attached selling prices to the mugs that were far higher than the buying prices chosen by randomly selected individuals who did not receive mugs. Evidently, being "endowed" with a mug greatly affected attitudes toward the mug, since individuals attached value not just to the end states (having versus not having a mug) but also to the transitions (receiving versus giving up a mug) (Kahneman, 2000).

The mugs experiments were important in part for ruling out a host of potential alternative explanations for the observed endowment effect. In addition to allocating mugs, the experimenters in the mugs study allocated tokens with preassigned cash values (the amounts for which subjects could redeem the tokens at the end of the study). Trading in tokens followed precisely the predictions of traditional economic theory; exactly half of the tokens changed hands, as theory would predict in light of the random assignment of tokens, and thus neither transaction costs nor other general trading barriers could explain the behavior observed in the case of the mugs.

The endowment effect has many important implications for legal design generally, as discussed by Jolls (2007b). This essay overviews a key set of implications in the specific domain of employment law.

THE ENDOWMENT EFFECT AND EMPLOYMENT MANDATES

A frequent claim in law and economics analysis of employment law is that the imposition of mandatory employment terms will tend to reduce employment levels by operating as a tax on their transaction (e.g., Posner, 1998). On this view, because the parties did not bargain for the term in question when left to their own devices, the cost of the term must exceed its benefit (for otherwise they would have agreed to it on their own). Thus, for example, if a particular employment benefit is worth $100 per year to employees and costs the employer only $90 to provide, a mandate should not be necessary; but if we do not observe the parties agreeing to the benefit on their own, then the cost must exceed $100. Imposing a mandatory term in these circumstances will operate as a tax on the parties, causing the wage to fall by somewhere between the benefit and the cost of the term and causing the employment level to fall (Summers, 1989).

The endowment effect calls this account into question. As described above, the endowment effect implies that people are often less willing to sell entitlements that are given to them than to buy entitlements that they do not already possess; if given a mug, they will not sell it for $\$X$, but if not given a mug, they will not buy one for that price. Thus, the fact that employees choose not to purchase a particular workplace benefit if they are not granted an entitlement to it does not imply that they would want to sell the entitlement (if they could) once it has been granted. The corollary of this observation is that imposing a mandatory term may have different effects than the standard analysis predicts (Craswell, 1991). In supply-and-demand terms, imagine a labor supply curve reflecting willingness to work at different wage levels given provision of the benefit; the consequence of the endowment effect may be that this curve is shifted to the right once the mandate is imposed, and this move may more than compensate for the backward shift in the employer's labor demand curve as a result of the mandate. In such a case, the wages of the affected worker will fall by as much as or more than the cost of the benefit.

Empirical evidence provides support for the endowment effect analysis of employment mandates. The seminal study in this area is that of Gruber (1994), which examined the effects of imposing mandatory coverage of childbirth expenses in employer-provided insurance policies. Imposition of the mandatory health-insurance term—which represented a substantial departure from the usual contractual arrangements prior to the mandate—caused the wages

of affected workers (most prominently, married women of childbearing age) to fall by at least the cost of the mandated coverage according to most of the author's estimates. The study also found that the hours of employment of these workers were either unchanged or slightly higher with the mandate and that their probability of being employed was either unchanged or slightly lower. In sum, "the findings consistently suggest shifting of the costs of the mandates on the order of 100 percent, with little effect on net labor input" (Gruber, 1994, p. 623).

These findings are difficult to reconcile with the Posnerian account, which predicts a fall in wages less than the cost of the benefit. (If the wage were going to adjust by the full cost of the benefit, then some substantial fraction of employers should have offered the benefit even prior to the mandate.) Of course, if the Posnerian account is modified to incorporate a conventional market failure such as adverse selection, then Gruber's findings may be explained without reference to the endowment effect, as Gruber notes.

Several caveats to the endowment effect analysis bear emphasis. First, while the endowment effect is consistent with complete or more-than-complete adjustment of the wage, it is also possible to have less-than-complete adjustment of the wage in the presence of the endowment effect. Perhaps workers are not any more willing to supply labor in exchange for a given wage plus the benefit in question once they have an entitlement to the benefit; it may be just that they would be even less willing to supply labor in the absence of the benefit.

The second qualification is that the endowment effect may not operate in contexts in which the beneficiaries of a mandatory term must give up a preexisting level of income, since they may be highly averse to such a loss (see, e.g., Kahneman and Tversky, 1979). This qualification is potentially important, as employees may face a financial loss relative to some preexisting expectation when a new benefit is mandated.

The final qualification is that the analysis offered here is purely positive, concerned with the effects of imposing a mandatory employment term. The endowment effect does not necessarily imply that, from a normative perspective, such terms are desirable; they may be efficient, in the sense that they would not be undone (if they could be) once imposed, but the situation without such terms is also efficient, for the same reasons given by the Posnerian account, and there is no obvious means by which the two situations can be compared.[18]

Conclusion

A wide range of employment law rules—from wage payment law to pension regulation to minimum wages to prohibitions on employment discrimination to mandated health insurance and workplace leave—are illuminated by consideration of bounded willpower, bounded self-interest, and bounded rationality. The effects of employment law turn in significant part on how employees and employers act in response to this law and thus it is not at all surprising that behavioral economics can help both to understand and in some cases to improve employment law.

Notes

The sections "Bounded Self-Interest and Minimum Wage Regulation" and "Bounded Rationality, Employment Discrimination Law, and Employment Mandates" draw from my previous work, while the balance of the essay is original. Thanks to Bruce Ackerman, Alan Schwartz, and Princeton conference participants for helpful comments and to Shuky Ehrenberg and Daniel Klaff for excellent research assistance.

1. As Bernheim (1995) noted, it is not clear that respondents understand "annual household income" the way economists do, but the substantial gap between the answers to the two very similarly worded questions seems hard to explain as an artifact of such potential limits on respondents' understanding.

2. The conclusion about Pareto inferiority assumes that when the individual is unconstrained, the individual will choose a constant consumption level across periods. To understand this assumption, note that the individual's multiple temporal selves may be viewed as players in a noncooperative game (Laibson, 1996). The game between the temporal selves will often have multiple equilibria; however, plans calling for consumption at a constant rate each period can reasonably be thought of as focal points.

Note that throughout the analysis in this section, individuals are assumed to be aware of their bounded willpower and its consequences for behavior in each period—the assumption utilized in most of the economics literature on bounded willpower. In the terminology of O'Donoghue and Rabin (1999), individuals are assumed to be "sophisticates" rather than "naifs." While Akerlof (1991) claimed that he was a "naif" in his decision making about mailing a package of Joe Stiglitz's clothes from India back to Stiglitz, hopefully he will forgive readers who feel somewhat skeptical about this claim.

3. See http://www.dol.gov/ebsa/FAQs/faq_compliance_pension.html

4. One might wonder why ERISA is needed at all to achieve the flexible regime just described; why could private contracting not achieve precisely the same arrangement? The historical difficulty with contracting in the pension context—a difficulty that led to the enactment of ERISA—was that employers and employees often held widely differing interpretations of private "pension contracts," leading to great uncertainty over the meaning of such contracts

(Nader and Blackwell, 1973). ERISA standardized pension arrangements—along with offering substantial tax advantages that undoubtedly further encouraged the growth of pension plans. Note that the tax subsidy for pensions encourages retirement saving among all individuals—whether or not they have bounded willpower—simply by lowering its cost.

5. The section "Bounded Self-Interest and Minimum Wage Regulation" is an abridged version of Jolls (2002). Here and below, when I use material from my previous work, the discussion will, as a result, not incorporate newer sources that have appeared since the original publication date, although, of course, if any development had altered the previously published analysis in any significant way, the analysis would have been updated to reflect the development.

6. See Kaufman's (1999) discussion of wage determination for a useful summary.

7. See Jolls (2002), for further detail on the empirical literature in this area.

8. To provide the strongest possible test of the fairness hypothesis, Fehr, Kirchsteiger, and Riedl used the labels "buyer" and "seller" for subjects assigned (respectively) to the "employer" and "employee" groups. This is likely to provide the strongest possible test because fairness considerations seem more likely to be present in employer-employee relationships, which usually involve social interaction, than in the relationship between the buyer and the seller of a type of good other than labor. If fairness is important even when the labels of "employer" and "employee" are not used, then it is even more likely to be important (or at a minimum is no less likely to be important) in a setting in which those labels are used. Consistent with this conclusion, a subsequent article by Fehr et al. (1998) that replicated the original test using employer and employee labels found strong effects of fairness. For expositional ease, I use the "employer" and "employee" labels.

9. This emphasis marks a contrast with that of Akerlof's original fair-effort work (1982), which took as its motivation a situation in which employees—young women in the first part of the twentieth century—were not difficult to monitor (indeed their output was known with exactitude) but were difficult to punish because their attachment to the labor force was quite limited (because most left the job within a short time to marry).

10. Here and below, see Jolls (2002), for citations to specific statutory sections and court cases.

11. *Massachusetts Labor Bulletin*, *13*, 1 (1900).

12. Again, see Jolls (2002), for detail on the legal provisions.

13. The material describing the nature of judgment errors is a slightly modified version of material from the introduction and part I of Jolls and Sunstein (2006).

14. Kahneman and Frederick (2002) provided a succinct description of the study and its results.

15. For further discussion, as well as qualifications, see Jolls (2007a).

16. See Jolls (2007a), for a discussion of related studies.

17. The material on employment mandates is a modified version of material from part II.D of Jolls, Sunstein, and Thaler (1998).

18. The endowment effect is naturally viewed as a species of a broader "status quo effect," under which the status quo, whatever shape it takes, tends to stick. As noted earlier, the status quo effect has been extensively studied in the context of retirement saving.

References

Akerlof, G. A. (1982). Labor contracts as partial gift exchange. *Quarterly Journal of Economics*, *97*, 543–569.

———. (1991). Procrastination and obedience. *American Economic Review*, *81*, 1–19.

Akerlof, G. A., and Yellen, J. L. (1990). The fair wage-effort hypothesis and unemployment. *Quarterly Journal of Economics*, *105*, 255–283.

Amador, M., Werning, I., and Angeletos, G. (2006). Commitment vs. flexibility. *Econometrica*, *74*, 365–396.

Bankman, J. (1988). Tax policy and retirement income: Are pension plan anti-discrimination provisions desirable? *University of Chicago Law Review*, *55*, 790–835.

Benartzi, S., and Thaler, R. H. (2007). Heuristics and biases in retirement savings behavior. *Journal of Economic Perspectives*, *21*(3), 81–104.

Bernheim, B. D. (1995). Do households appreciate their financial vulnerabilities? An analysis of actions, perceptions, and public policy. In *Tax policy for economic growth* (pp. 3–30). Washington, DC: ACCF.

Beshears, J., Choi, J. J., Laibson, D., and Madrian, B. C. (2005). Early decisions: A regulatory framework. *Swedish Economic Policy Review*, *12*, 41–60.

Bewley, T. F. (1999). *Why wages don't fall during a recession*. Cambridge, MA: Harvard University Press.

Blinder, A. S., and Choi, D. H. (1990). A shred of evidence on theories of wage stickiness. *Quarterly Journal of Economics*, *105*, 1003–1015.

Campbell, C. M., III, and Kamlani, K. S. (1997). The reasons for wage rigidity: Evidence from a survey of firms. *Quarterly Journal of Economics*, *112*, 759–789.

Chaiken, S., and Trope, Y. (Eds.) (1999). *Dual-process theories in social psychology*. New York: Guildford.

Craswell, R. (1991). Passing on the costs of legal rules: Efficiency and distribution in buyer-seller relationships. *Stanford Law Review*, *43*, 361–398.

Dasgupta, N., and Asgari, S. (2004). Seeing is believing: Exposure to counterstereotypic women leaders and its effect on the malleability of automatic gender stereotyping. *Journal of Experimental Social Psychology*, *40*, 642–658.

Eaton, L. (1998, July 26). Show nanny the money: As economy booms, pay rises for child-care workers. *New York Times*, p. 27.

Fehr, E., Kirchler, E., Weichbold, A., and Gächter, S. (1998). When social norms overpower competition: Gift exchange in experimental labor markets. *Journal of Labor Economics*, *16*, 324–351.

Fehr, E., Kirchsteiger, G., and Riedl, A. (1993). Does fairness prevent market clearing? An experimental investigation. *Quarterly Journal of Economics*, *108*, 437–459.

Feldstein, M., and Liebman, J. B. (2002). Social security. In A. J. Auerbach and M. Feldstein (Eds.), *Handbook of public economics* (Vol. 4, pp. 2245–2324). Amsterdam: Elsevier.

Frank, R. H., and Hutchens, R. M. (1993). Wages, seniority, and the demand for rising consumption profiles. *Journal of Economic Behavior and Organization*, *21*, 251–276.

Frederick, S., Loewenstein, G., and O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature*, *40*, 351–401.

Greenwald, A. G., McGhee, D. E., and Schwartz, J.L.K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*, 1464–1480.

Gruber, J. (1994). The incidence of mandated maternity benefits. *American Economic Review*, *84*, 622–641.

Imrohoroglu, A., Imrohoroglu, S., and Joines, D. H. (2003). Time-inconsistent preferences and Social Security. *Quarterly Journal of Economics*, 118, 745–784.

Jolls, C. (1996). Hands-tying and the Age Discrimination in Employment Act. *Texas Law Review*, *74*, 1813–1846.

———. (2002). Fairness, minimum wage law, and employee benefits. *New York University Law Review*, *77*, 47–70.

———. (2007a). Antidiscrimination law's effects on implicit bias. Retrieved from http://www.law.yale.edu/documents/pdf/Antidiscrimination Laws Effects.pdf (Previously published in M. Gulati and M. Yelnosky [Eds.], *Behavioral analyses of workplace discrimination* [Vol. 3 of *NYU selected essays on labor and employment law*, pp. 69–102]. The Netherlands: Kluwer Law International).

———. (2007b). Behavioral law and economics. Retrieved from http:// www.law.yale.edu/documents/pdf/Faculty/Jolls BehavioralLawand Economics.pdf (Previously published in P. Diamond and H. Vartiainen [Eds.], *Behavioral economics and its applications* [pp. 115–156]. Princeton, N.J.: Princeton University Press).

Jolls, C., and Sunstein, C. R. (2006). The law of implicit bias. *California Law Review*, *94*, 969–996.

Jolls, C., Sunstein, C. R., and Thaler, R. (1998). A behavioral approach to law and economics. *Stanford Law Review*, *50*, 1471–1550.

Kahneman, D. (2000). New challenges to the rationality assumption. In D. Kahneman and A. Tversky (Eds.), *Choices, values, and frames* (pp. 758–774). Cambridge: Cambridge University Press and New York: Russell Sage Foundation.

Kahneman, D., and Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, and D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). Cambridge: Cambridge University Press.

Kahneman, D., Knetsch, J. L., and Thaler, R. (1986). Fairness as a constraint on profit-seeking: Entitlements in the market. *American Economic Review*, *76*, 728–741.

———. (1990). Experimental tests of the endowment effect and the Coase Theorem. *Journal of Political Economy*, *98*, 1325–1348.

Kahneman, D., and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–291.

Kaufman, B. E. (1999). Expanding the behavioral foundations of labor economics. *Industrial and Labor Relations Review*, *52*, 361–392.

Krieger, L. H., and Fiske, S. T. (2006). Behavioral realism in employment discrimination law: Implicit bias and disparate treatment. *California Law Review*, *94*, 997–1062.

Kumru, C. S., and Thanopoulos, A. C. (2008). Social security and self control preferences. *Journal of Economic Dynamics and Control*, *32*, 757–778.

Laibson, D. (1996). *Hyperbolic discount functions, undersaving, and savings policy.* NBER Working Paper No. 5635. National Bureau of Economic Research.

Laibson, D. I., Repetto, A., Tobacman, J., Hall, R. E., Gale, W. G., and Akerlof, G. A. (1998). Self-control and saving for retirement. *Brookings Papers on Economic Activity*, *1998*(1), 91–196.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., and Vishny, R. W. (1997). Trust in large organizations. *American Economic Review, 87*, 333–338.

Lazear, E. P. (1979). Why is there mandatory retirement? *Journal of Political Economy*, *87*, 1261–1284.

Levine, D. I. (1993). Fairness, markets, and ability to pay: Evidence from compensation executives. *American Economic Review*, *83*, 1241–1259.

Linder, M. (1987). Farm workers and the Fair Labor Standards Act: Racial discrimination in the New Deal. *Texas Law Review*, *65*, 1335–1393.

———. (1992). *Migrant workers and minimum wages: Regulating the exploitation of agricultural labor in the United States.* Boulder, CO: Westview Press.

Loewenstein, G., and Sicherman, N. (1991). Do workers prefer increasing wage profiles? *Journal of Labor Economics*, *9*, 67–84.

Lowery, B. S., Hardin, C. D., and Sinclair, S. (2001). Social influence effects on automatic racial prejudice. *Journal of Personality and Social Psychology*, *81*, 842–855.

Medoff, J. L., and Abraham, K. G. (1980). Experience, performance, and earnings. *Quarterly Journal of Economics*, *95*, 703–736.

———. (1981). Are those paid more really more productive? The case of experience, *Journal of Human Resources*, *16*, 186–216.

Nader, R., and Blackwell, K. (1973). *You and your pension*. New York: Grossman.

Neumark, D., and Stock, W. A. (1999). Age discrimination laws and labor market efficiency. *Journal of Political Economy*, *107*, 1081–1125.

Nosek, B. A., Banaji, M. R., and Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration website. *Group Dynamics: Theory, Research, and Practice*, *6*, 101–115.

O'Donoghue, T., and Rabin, M. (1999). Doing it now or later. *American Economic Review*, *89*, 103–124.

Phelps, E. S., and Pollak, R. A. (1968). On second-best national saving and game-equilibrium growth. *Review of Economic Studies*, *35*, 185–199.

Posner, R. A. (1998). *Economic analysis of law* (5th ed.). New York: Aspen Law and Business.

Richeson, J. A., and Ambady, N. (2003). Effects of situational power on automatic racial prejudice. *Journal of Experimental Social Psychology*, *39*, 177–183.

Slichter, S. H. (1929). The current labor policies of American industries. *Quarterly Journal of Economics*, *43*, 393–435.

Smith, P. R. (1999). Regulating paid household work: Class, gender, race, and agendas of reform. *American University Law Review*, *48*, 851–924.

———. (2000). Organizing the unorganizable: Private paid household workers and approaches to employee representation. *North Carolina Law Review*, *79*, 45–110.

Solow, R. M. (1990). *The labor market as a social institution*. Cambridge, MA: Blackwell.

Summers, L. H. (1989). Some simple economics of mandated benefits. *American Economic Review*, *79*, 177–183.

Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior and Organization*, *1*, 39–60.

———. (1981). Some empirical evidence on dynamic inconsistency. *Economics Letters*, *8*, 201–207.

———. (1990). Anomalies: Saving, fungibility, and mental accounts. *Journal of Economic Perspectives*, *4*(1), 193–205.

———. (1994). Psychology and savings policies. *American Economic Review*, *84*, 186–192.

———. (1996). Doing economics without *homo economicus*. In S. G. Medema and W. J. Samuels (Eds.), *Foundations of research in economics: How do economists do economics?* (pp. 227–237). Cheltenham, UK: Edward Elgar.

Thaler, R. H., and Benartzi, S. (2004). Save More Tomorrow: Using behavioral economics to increase employee saving. *Journal of Political Economy*, *112*, S164–S187.

Thaler, R. H., and Shefrin, H. M. (1981). An economic theory of self-control. *Journal of Political Economy*, *89*, 392–406.

Weiss, D. M. (1991). Paternalistic pension policy: Psychological evidence and economic theory. *University of Chicago Law Review*, *58*, 1275–1319.

Zelizer, V. A. (1994). *The social meaning of money*. New York: Basic Books.

CHAPTER 16

Decision Making and Policy in Contexts of Poverty

SENDHIL MULLAINATHAN

ELDAR SHAFIR

Policy thinking about poverty typically falls into two camps. Social scientists as well as regular folk regard the behaviors of the economically disadvantaged either as calculated adaptations to prevailing circumstances or as emanating from a unique "culture of poverty" that is rife with deviant values. The first view presumes that people are highly rational, that they hold coherent, well-informed, and justified beliefs and that they pursue their goals effectively with little error and with no need for help. The second perspective attributes to the poor a variety of psychological and attitudinal shortcomings—failings that render their views often misguided, their behaviors lacking, and their choices fallible, leaving them in need of paternalistic guidance.

Both camps are likely to capture some important elements some of the time. There are, no doubt, important circumstances where people—the poor included—are methodical and calculating, and other circumstances where they are fallible or misguided. But both fail to explain important phenomena. We propose an alternative perspective, one that is largely informed by recent behavioral research. According to this perspective, the behavioral patterns of the poor may be neither perfectly calculating nor especially deviant. Rather, the poor may exhibit fundamental attitudes and natural proclivities, including weaknesses and biases, that are similar to those of people from other walks of life. One important difference, however, is that in poverty there are narrow margins for error, so that the same behaviors often manifest themselves in more pronounced ways and can lead to worse outcomes (see Bertrand, Mullainathan, and Shafir, 2004, 2006). In fact, we argue, the same psychological proclivities in the context of poverty can ultimately yield particular behavioral patterns that are endemic to functioning under scarcity.

The "rational" view assumes the poor are doing as well as they can and ought to be left to their own devices, hopefully in a context that affords good options. The culture-of-poverty perspective is motivated by the impulse to change how the poor function. In contrast, the central gist of the behavioral perspective is that, much of the time, the poor are not functioning optimally, nor are they in need of behavioral change anymore than the rest of us. Instead, it is the interaction of fundamental behavioral proclivities and the context in which they function that produces the particularly destructive circumstances in which the poor often find themselves. According to this view, people who live in poverty are susceptible to many of the same impulses and idiosyncrasies as those who live in comfort, but whereas people who are better-off find themselves in the midst of a system—composed of consultants, reminders, cooperative employers, no-fee options, incentive awards, and automatic deposits—that is increasingly designed to facilitate their decisions and improve their outcomes, people who are less well-off typically find themselves without easy recourse to such aids and often confront obstacles—institutional, social, and psychological—that render their economic conduct all the more overwhelming and fallible.

In what follows, we will describe our work on a psychologically more realistic analysis of poverty. Our thinking has evolved over time. For the most part, it involves a direct application of the empirical research on judgment and decision making supplemented by lessons from social and cognitive psychology. Our approach was simple: psychology gives us insights into how people behave. How might these behaviors have different consequences when people are poor? For example, we know that self-control can be difficult. However, the same self-control failure (e.g., giving in to a temptation and buying something you should not) can have larger consequences for a poor person than for one for whom the expense is barely detectible (this discussion borrows heavily from Mullainathan

and Shafir, 2009). Such a perspective provides several ideas for interventions that might reduce the impact of poverty. Of course, insights generated by experimental research and empirical observation need to be carefully tested and evaluated before they are relied on to shape policy. Even when an intervention succeeds in producing some intended outcome, there is always the possibility that other, unforeseen patterns will emerge. Bearing that in mind, we propose some guidelines for thinking about the future design of financial services. Related concerns regarding regulation are addressed in Barr, Mullainathan, and Shafir (this volume).

Our latest thinking has been more radical, and we describe it briefly in the final section. There we will argue that poverty itself generates *specific* psychological responses. The lack of slack that characterizes poverty, we suggest, can itself affect the operation of the mind and the making of decisions. For example, when we are focused on making ends meet, we end up distracted by local concerns and calculations. We are less likely to weigh long-term consequences and exhibit forward-looking behaviors when we are threatened, challenged, and depleted.

This in turn suggests new approaches to policy making that more directly address instability, so that programs may realize greater promise. We describe possible strategies for redesigning existing programs that can foster stability and give people the financial and psychic steadiness they need to make better choices and build better lives. We will conclude with brief summary remarks and a discussion of future research directions.

The Behavioral Perspective

Context Dependence

Human behavior has proven to be heavily context dependent, a function of both the person and the situation. One of the major lessons of modern psychological research is the impressive power that the situation exerts, along with a persistent tendency to underestimate that power relative to the presumed influence of personality traits. Various studies have documented the stunning capacity of situational factors to influence behaviors that are typically seen as reflective of deep personal dispositions. For example, in his now-classic obedience studies, Milgram (1974) showed how decidedly mild situational pressures sufficed to generate persistent willingness on the part of regular people to administer what they believed to be grave levels of electric shock to innocent others. Along similar lines, Darley and Batson (1973) recruited seminary students to deliver a practice sermon on the parable of the Good Samaritan. While half the seminarians were told they were ahead of schedule, others were led to believe they were running late. On their way to give the talk, all participants passed an ostensibly injured man slumped in a doorway, coughing and groaning. Whereas the majority of those with time to spare stopped to help, a mere 10% of those who were running late stopped, the remaining 90% stepping over the victim and rushing along. These participants' ethical training and biblical scholarship notwithstanding, the contextual nuance of a minor time constraint proved decisive in the decision to stop and help a suffering man.

Construal

A simple but fundamental tension between classical economic analyses and modern psychological research is captured by the role of "construal." Agents in classical economic analyses are presumed to choose between options in the world, objectively represented. People, however, do not respond directly to objective circumstances; rather, stimuli are mentally construed, interpreted, understood (or misunderstood), and then acted upon. Behavior is directed not toward actual states of the world, but rather toward our mental representation of those states, and mental representations do not bear a one-to-one relationship to the thing they represent, nor do they necessarily constitute faithful renditions of actual circumstances. As a result, many interventions can fail because of the way in which they are construed by the targeted group. Some, for example, may interpret a well-intentioned intervention "as an insulting and stigmatizing exercise in co-option and paternalism" (Ross and Nisbett, 1991) or as an unintended indication of what the desired, or expected, behavior might be or of what it might be worth. Thus, people who are rewarded for a behavior they find interesting and enjoyable can come to attribute their interest in the behavior to the reward and, consequently, come to view the behavior as less appealing (Lepper, Greene, and Nisbett, 1973). In one classic study, children who were offered a "good player award" to play with magic markers, which they had previously done with great relish in the absence of extrinsic rewards, subsequently showed little interest in the markers when these were introduced as an unrewarded classroom activity (in contrast with kids who were not rewarded and did not lose interest).

Mental Accounting and Finances

One domain that is of great relevance to our present topic and where construal can prove of great consequence is that of mental accounting. Research on mental accounting documents the variety of ways in

which the assumption of the fungibility of money fails, leading people to view cash, credit, and debit differently depending on which "mental account" the money is perceived to be in. People's representation of money systematically departs from what is commonly assumed in economics. According to the fungibility assumption, which plays a central role in theories of consumption and savings, "money has no labels"; all components of a person's wealth can be collapsed into a single sum. Contrary to this assumption, people appear to compartmentalize wealth and spending into distinct budget categories, such as savings, rent, and entertainment, and into separate mental accounts, such as current income, assets, and future income (Thaler, 1985, 1992). These mental accounting schemes lead to differential marginal propensities to consume (MPC) from one's current income (where MPC is high), current assets (where MPC is intermediate), and future income (where MPC is low). Consumption thus ends up being overly dependent on the mental accounting of current wealth, so that, for example, people find themselves willing to save and borrow (at a higher interest rate) at the same time (Ausubel, 1991).

There are a variety of other experimental findings that are relevant to a deep understanding of financial behaviors but which are beyond the purview of the present brief exposition. To list just a few, people are loss averse (the loss of utility associated with giving up a good is greater than the utility associated with obtaining it; Tversky and Kahneman, 1991), and loss aversion yields "endowment effects," wherein the mere possession of a good can lead to higher valuation of it than if it were not in one's possession (Kahneman, Knetsch, and Thaler, 1990). This, in turn, leads to a general reluctance to depart from the status quo, because the disadvantages of departing from it tend to loom larger than the advantages of the alternatives (Knetsch, 1989; Samuelson and Zeckhauser, 1988). People often also fail to ignore sunk costs (Arkes and Blumer, 1985), fail appropriately to consider opportunity costs (Camerer et al., 1997), and show money illusion, wherein the nominal worth of money interferes with a representation of its real worth (Shafir, Diamond, and Tversky, 1997). Furthermore, people often prove weak at predicting their future tastes or at learning from past experience (Kahneman, 1994), their intertemporal choices exhibit poor planning (Buehler, Griffin, and Ross, 1994), and they exhibit high discount rates for future, as opposed to present, outcomes, yielding dynamically inconsistent preferences (Loewenstein and Prelec, 1992; Loewenstein and Thaler, 1989).

An understanding of such proclivities may be harnessed to help make sense of behaviors that otherwise appear perplexing and may help produce more desirable behaviors and better policy outcomes. For example, due to faulty planning and procrastination, numerous studies of middle-class workers have shown that saving works best as a default. Thus, participation in 401(k) plans is significantly higher when employers offer automatic enrollment (Madrian and Shea, 2001). And because participants tend to retain the default contribution rates, savings can be increased as a result of agreeing to increased default deductions from future raises (Benartzi and Thaler; 2004). As we discuss below, the poor tend to have little recourse to just this kinds of default saving programs, but the general notion that the context can be designed so as to ameliorate outcomes is a central and important one.

Channel Factors

As it turns out, the pressures exerted by apparently trivial situational factors can create restraining forces that are hard to overcome or can promote inducing forces that can be harnessed to great effect. What is particularly impressive is the fluidity with which construal occurs and the sweeping picture it imposes. Alongside the remarkably powerful impact of context emerges a profound underappreciation of its effects. When interpreting others' behavior, there is a tendency to exhibit the *fundamental attribution error*, wherein the influence of internal, personal attributes is overweighted and the influence of external, situational forces is underappreciated. As explained by Ross and Nisbett (1991), where standard intuition would hold the primary cause of a problem to be human frailty, or the particular weakness of a group of individuals, the social psychologist would often look to situational barriers and to ways to overcome them.

The basic insights outlined above have important corollaries for our present concerns. For one, they suggest that the same tendencies and weaknesses will express themselves differentially in diverse circumstances. For example, the tendency to avoid action and resort to the status quo will lead to inferior outcomes when the context is structured so that the most beneficial outcomes require action, and it will lead to more desirable outcomes whenever the default is set naturally to produce them. Similarly, a person who is well-off and fails to formulate a farsighted plan may find himself with a more modest though still comfortable nest egg upon retirement, whereas a poor person who exhibits such failures may end up with too little cash to pay a phone bill, accrue large fines for reconnection, experience increased inability to pay bills, and descend further into poverty.

Identity

Recent research has highlighted the role of identity salience in decision making (see, e.g., LeBoeuf,

Shafir, and Bayuk, 2010, and references therein). People derive their identity in large part from the social groups to which they belong (Turner, 1987). A person may alternate among different identities—she might think of herself primarily as a mother when in the company of her children, but see herself primarily as a professional while at work. The list of possible identities is extensive, with some identities (e.g., mother) likely to conjure up strikingly different values and ideals from others (e.g., CEO).

Of particular relevance here might be the natural salience of a "poor, incapable, untrustworthy" identity, that is likely to loom in the background of any potential transaction, and could have substantially detrimental effects. Several studies have confirmed the notion of "stereotype threat" (Steele, 1997; Steele and Aronson, 1995), according to which a prevalent stereotype about a group creates a burden on group members that acts as a threat. The threat arises whenever stigmatized individuals' behavior runs the risk of substantiating the stereotype, and this threat can distort or disrupt the performance of those individuals. In one study, for example, Asian women whose race (stereotypically strong in math) was made salient performed significantly better on a tough mathematics exam than when their gender (stereotypically weak in math) was rendered salient (Shih, Pitinski, and Ambady, 1999). Several studies have shown similar effects with African Americans, and some have replicated these effects on people from low socioeconomic backgrounds. When students from a low socioeconomic status (SES) are subjected to doubts about their intellectual ability that are similar in kind to those experienced by African Americans, the threat has similarly disruptive effects. In one study, low-SES students performed worse than high-SES students when the test was presented as a measure of intellectual ability; however, the low-SES students' performances matched those of the high-SES students when the test was not presented as measuring intellectual ability (Croizet and Claire, 1998).

Similar phenomena will likely be observed in other behavioral domains; for example, where stereotypes involving intellectual and professional ability might interfere with a person's willingness to, say, interact with a bank. Adkins and Ozanne (2005) discussed the impact of a low-literacy identity on consumers' behavior and argued that when low-literacy consumers accept the low-literacy stigma, they perceive market interactions as riskier, engage in less extended problem solving, limit their social exposure, and experience greater stress.

Along similar lines, some have suggested that one reason for the relative success of Earned Income Tax Credit (EITC) is that it explicitly appeals to people's identity as taxpayers, rather than poor. In fact, specific personality traits, such as regulatory orientation (promotion versus prevention), may also fit certain decision-making contexts better than others. Research by Higgins and colleagues (Higgins, 2000; Higgins et al., 2003) has shown that people value items more when these were chosen using a strategy that fit their orientation rather than a strategy that did not fit. (Related discussion of the role of identity and construal is provided by Briley and Aaker, 2006a, 2006b.)

All of the above suggests that when it comes to bank accounts and other services intended for the poor, the government and banks may want to promote such services to those identities—head of family, working taxpayer—that might trigger a more positive response in the intended recipients.

In what follows, we will examine some specific implications of the behavioral perspective for the financial lives of the poor. At the individual level, what does it entail for choices about savings and borrowing, and at the institutional level, what does this perspective say about how financial services ought to be designed? (More about the regulatory implications of this perspective is explored in Barr, Mullainathan, and Shafir, in this volume.) Individual psychology is relevant at each of these levels. It affects the choices and actions that compound to generate a pattern of saving and borrowing. It affects how individuals respond to various features of financial products, from pricing to transaction costs to its intertemporal consequences. And it can provide a different perspective on the channels by which financial services can affect behavior. Finally, since a behavioral analysis generates deviations from the traditional economic model, it also provides different rationales and guidance for regulation, which are sometimes in tension with the traditional assumptions that guide consumer protection.

Institutional Financial Access for the Poor

The Role of Financial Access

Financial services may provide an important pathway out of poverty. They facilitate savings to mitigate shocks or promote asset development, and they facilitate borrowing to purchase durables or help weather tough times. In short, they allow individuals to smooth consumption and invest. (For more on financial instruments used by low-income Americans, see Barr, 2004, and Barr et al., in this volume.) Improvement of financial services then provides two key advantages. First, it lowers the costs that individuals who were already accessing such services may need to pay. For

example, they may now be able to use a credit card rather than resort to more expensive payday lenders. Second, individuals who beforehand did not have access to such services may now get the direct benefit of access. For example, the ability to borrow may allow individuals to smooth shocks (such as health shocks).

Some Features of Financial Access

Our perspective highlights the importance of contextual nuance and focuses attention on the nature of circumstances that emanate from the interaction between behavioral tendencies and contextual structure. Here we briefly consider some simple contextual features pertinent to financial access.

INSTITUTIONS SHAPE DEFAULTS

It is well established that defaults can have a profound influence on the outcomes of individual choices. Data available on decisions ranging from retirement savings and portfolio choices to the decision to be a willing organ donor illustrate the substantial increase in market share of default options (Johnson and Goldstein, 2003; Johnson et al., 1993). This is likely to prove of great importance for the design of financial services, which often shape default financial behaviors.

Consider, for example, two individuals with no access to credit cards: one has her paycheck directly deposited into a savings account, and the other does not. Whereas cash is not readily available to the first person, who needs to take active steps to withdraw it, cash is immediately available to the second, who must take active measures to save it. The greater tendency to spend cash found in the wallet than if it were in the bank (Thaler, 1999) suggests that the first, banked person will spend less on impulse and save more easily than the person who is unbanked. Holding risk and savings-related propensities constant, the first is likely to end up a more active and efficient saver than the second.

INSTITUTIONS SHAPE BEHAVIOR

Many low-income families are, in fact, savers, whether or not they resort to banks (Berry, 2004). Without the help of a financial institution, however, their savings are at greater risk (including theft, impulse spending, and access by household members), will grow more slowly, and may not be readily available to support access to reasonably priced credit in times of need. Institutions provide safety and control. In this sense, an institutional context may be even more critical for the poor than for the comfortable. In circumstances of dearth, temptation, distraction, and difficult management and control, those savers who are unbanked are likely to find it all the more difficult to succeed on the path to long-term prosperity.

In fact, a recent survey conducted by the American Payroll Association shows that "American employees are gaining confidence in direct deposit as a reliable method of payment that gives them greater control over their finances, and that employers are recognizing direct deposit as a low-cost employee benefit that can also save payroll processing time and money."[1] The employers of the poor, in contrast, often do not require nor propose electronic salary payments. Instead, they prefer not to offer direct deposit to hourly/nonexempt employees, temporary or seasonal employees, part-timers, union employees, and employees in remote locations, all categories that correlate with being low paid. The most frequently stated reasons for not offering direct deposit to these employees include the lack of processing time to meet standard industry (Automatic Clearing House) requirements, high turnover, and union contract restrictions. All this creates a clearly missed opportunity to offer favorable defaults to needy individuals, whose de facto default consists of taking a check, often after hours, to a place, often inconvenient, where it can be cashed for a hefty fee.

INSTITUTIONS PROVIDE IMPLICIT PLANNING

As it turns out, a variety of institutions provide implicit planning, often in ways that address potential behavioral weaknesses. Credit-card companies send customers timely reminders of due payments, and clients can elect to have their utility bills automatically charged, allowing them to avoid late fees if occasionally they do not get around to paying in time. The low-income client, on the other hand, without the credit card, the automatic billing, or the web-based reminders, risks missed payments, (high) late fees, disconnected utilities (accompanied by high reconnection charges), etc.

Interestingly, context can also be detrimental by providing debt too easily. Temporal discounting in general, and present bias in particular, can be exploited to make immediate cash more attractive than the future costs of borrowing appear menacing. Whenever this happens, the increased availability of debt could lower the well-being particularly of the poor since overspending by the poor may entail subsequent cutbacks in far more essential consumption than overspending by the rich.

One fundamental lesson of the behavioral analysis for policy makers and regulators is a new appreciation for the impact and responsibility of financial institutions. These should not simply be viewed from a

financial cost-saving perspective but should instead be understood to affect the lives of people by easing their planning, facilitating their desired actions, or enabling their resistance to temptation. Such effects, furthermore, may have substantially different implications for those who are wealthier, who get professional help, and, at the same time, can afford to err or be tempted, as opposed to the poor, who resort to fewer professionals and may pay dearly even for infrequent giving-in to temptations or minor mistakes.

These considerations form part of a more general view of why financial institutions can be so important in the lives of the poor. Access to financial institutions allows people to improve their planning by keeping money out of temptation's way. In some cases (such as direct deposits and automatic deductions), one may not even notice the moment that the money "arrived" into savings or was invested into the long term. The recourse to financial institutions provides the opportunity to make infrequent, carefully considered financial accounting decisions, which can prove resistant to intuitive error or to momentary mental-accounting impulses. In this sense, improving financial institutions can have a disproportionate impact on the lives of the poor. Moving from a payday lender and check casher to a bank with direct deposit and payroll deduction can have benefits that far exceed the transactional costs saved. (For further discussion and examples of savings instruments aligned with behavioral principles, see Tufano, 2009.)

Some Noninstitutional Aspects of the Financial Lives of the Poor

Aided by the insights above, we will aim to further understand the interactions of the poor with specific financial institutions. To begin, we will discuss three stylized facts about the financial lives of the poor that are noninstitutional but that we think are especially important to the behavioral perspective. These stylized facts are not necessarily psychological (two of them have very straightforward economic interpretations). Rather, they are facts that may render the impact of the relevant psychology particularly interesting and consequential.

Lack of Financial Slack

Although it is hard to define precisely in an economic model, the notion of "economic slack" is central to the lives of the poor. We define *slack* as the ease with which one can cut back consumption to satisfy an unexpected need. Under this definition, the poor appear to have less economic slack than the rich. Whereas a rich person can often cut back on (by their own admission) more frivolous spending, a poor person faced with a financially demanding situation is forced to cut back on essential expenses. There are two ways to understand this mechanism. The first, more traditional, vehicle is via diminishing returns. If both rich and poor face equivalent shocks and cut back on consumption by the same amount, the rich person will be cutting back on lower-marginal-utility consumption. The second, more psychological, vehicle concerns temptations. If the incidence of temptation spending decreases with wealth, the rich will be cutting back on precisely those goods that are less valuable from the point of view of past or future selves.

This analysis abstracts from the role of savings. One might argue that the poor, exactly because they face a more volatile environment, would put aside enough buffer-stock savings to handle the excess volatility. This would mean that a comparable size shock should be less likely to result in a poor person running out of savings. While plausible, we ignore this factor in the following conceptualization because a large amount of data shows that poorer families tend to have negligible liquid savings. The lack of buffer-stock savings is, we feel, one of the more interesting puzzles to understand in the financial lives of the poor; we return to this issue briefly in "Some Economic Behaviors of the Poor," below.

A lack of financial slack is particularly consequential when one considers the type of expenditures the poor might be forced to cut back on. One common finding in the literature is the frequent occurrence of late payments and phone and gas disconnections (and ensuing costly reconnections). Edin and Lein (1997) estimated that nearly 5% of annual income was spent on such reconnections. Many financial services impose late payments. These range from the expected (credit-card bills) to the unexpected (rent-to-own stores that penalize individuals for missing a payment by repossessing the item, thereby forcing a loss of all payments so far). Landlords can impose late fees. All sorts of bills, from utility to medical bills, usually have steep late payments. The key observation about fees is that they are usually disproportionate. For example, a 5% late fee for a monthly bill is effectively a 100% APR on a loan. In other words, if the poor cut back by skipping a bill payment, they are effectively borrowing at very high rates.

High-interest-rate borrowing may be the *least* costly consequence. In fact, what makes the lack of financial slack particularly onerous are the indirect, but linked, consequences. Consider a household that has had their phone disconnected. They now face several

difficult consequences. First, they need to make a large lump-sum payment to get their phone reconnected. Acquiring this large lump-sum poses extra difficulties to an already stretched budget. Second, and more important, the lack of a phone could have other consequences for their lives. For example, if they happen to be unemployed (not unlikely for a household that was unable to pay its phone bills), they are now far less effective job searchers. Even if they are employed, the employer may not be able to reach the home in case shifts change and they are needed at work, making them a less valuable employee. In other words, one action—paying the phone bill late—can have dynamic consequences, amplifying the initial cost and further depressing income. Low-income households struggling with the chronic lack of slack that comes with being low-income are thus always at risk of becoming ever more destitute.

There are profound consequences to being on the edge of further destitution. The first is that any failures to plan well can have quite severe consequences. A rich person who fails to plan, or plans poorly, will simply cut back on frivolous expenditures. A poorer individual may face a domino effect of consequences that can amplify an otherwise small misplanning step. The lack of slack means that the poor must walk a planning tightrope. They must in effect be superplanners in less conducive and less helpful surroundings lest they slip deeper into poverty.

A second consequence is empirically easier to identify. It means that the poor will sometimes exhibit a willingness to borrow at very high rates. The individual who is facing the prospect of having his phone shut off, with a hefty late fee to turn it on again and the assorted difficulties that arise from a lack of phone service, may well be willing to borrow at high rates to avoid this from happening or to get the phone reconnected if it has already happened. In fact, not only are low-income individuals willing to borrow at very high rates, it may be rational for them to do so. The desire to borrow at high rates is interesting because it can easily be confused for myopia, whereas in some contexts it can constitute a perfectly rational, even if undesirable, response. This is also relevant to payday loans, an issue we return to below.

Small to Big Transformations

One of the fundamental services that financial institutions provide is to allow for the gradual transformation of small amounts of cash, which are easier to come by, into larger lump sums, which can be hard to attain. As Rutherford (2001) explains, individuals often need to transform small cash amounts into "usably large" amounts. Such transformations often prove essential for the poor because of the nature of their cash inflows and needs. The urban poor typically deal with cash inflows in relatively small amounts. Having received weekly or biweekly paychecks, after the necessary rent, utility, and other bills, they are typically left with only small amounts of cash on hand. Yet, many of the durables they might wish to purchase—washing machines, cars, televisions—will require more than what they have left at any point in time. Consequently, the poor will need to transform small amounts into usably large sums.

According to traditional economic theory, such transformation is straightforward: individuals would simply save the cash they come by until they have accumulated enough. Alternatively, if credit is available, individuals could borrow against future income streams to finance the transformation. Whether debt or savings are used depends on the flow value of the durable to be purchased relative to the interest rate on debt. Of course, because the poor often do not have access to credit, they would need to save their way up.

The psychology of planning and self-control suggests that such savings may be more difficult than traditional theory is prone to assume. An individual saving to buy a durable over a long period of time would have large amounts of cash continuously accessible. And accessible cash can be extremely tempting, often leading it to be spent on things that are mostly valued at the moment of spending. As such, temporal inconsistency and self-control problems make savings a poor vehicle on which to rely for small-to-big cash transformations. They turn savings accounts into highly leaky budgets.

Many institutions that are popular among the poor, and which may otherwise appear as less than perfectly rational solutions, can be understood as alternative methods for making small-to-big transformation more feasible in a world of imperfect planning and limited control. First, consider the purchase of lottery tickets, which, as many have noted, the poor are especially likely to partake in (Blalock, Just, and Simon, 2007; Kearny, 2005). What is particularly interesting is the type of lottery ticket they typically buy—namely, for modest lotteries with maximum payoffs of roughly $200 to $500. If the poor are "buying dreams" through these lottery tickets, these are quite modest dreams. Such small maximum payoffs are more consistent with lottery tickets as a vehicle for small-to-big conversion. An individual who struggles to save up to buy a $400 item, for example, would find it easy to buy a lottery ticket periodically. The recurring costs are the "deposits," which eventually lead to a win and the ability to buy the expensive

item with the winnings. Notice the advantages of this over the typical savings account. No money accumulates that can be dipped into in the face of recurring temptation, vis-à-vis one's needs, or those of family and friends. The individual loses his outlay until he (effectively) wins the desired item, the lottery ticket essentially serving as a commitment device, albeit an expensive one.

Notice that this explanation is very similar to a self-control explanation for the prevalence of rotating savings and credit associations (ROSCAs) in developing countries (Basu, 2008). In a typical ROSCA, each participant contributes a fixed amount each week or month, with one participant taking the entire pot. The winner is determined by lottery or by bidding, with each participant eligible to win once throughout the ROSCA. This is much like a lottery ticket except that one is guaranteed to win 1 in N times. Both of these institutions reinforce the view that a bigger lump of money is worth more to the poor than many small amounts.

Perhaps most telling is the prevalence of layaway plans. In a typical layaway program, an individual picks a particular durable he would like, for example, a washing machine. He then opens a layaway account, to which he deposits money, with a regularity of payments that depends on the particular store. Once the client has accumulated enough, he is given the durable. This is quite similar to the SEED commitment savings product offered to clients of a Philippine bank by Ashraf, Karlan, and Yin (2006). Some stores offer a price lock-in feature, so that prospective buyers are guaranteed the initial posted price, but many others do not. Individuals who do not save enough to buy the item often forfeit their cash. It appears that the primary benefit of the layaway account is its illiquidity.

The popularity of layaways emphasizes the difficulty that simple myopia models face in explaining the behavior of the poor. In resorting to such arrangements, the poor are showing remarkable farsightedness. They are opting to save, without interest, in order to purchase a durable good, which they do not even get to enjoy as they save up to buy it. As with other examples in this section, there is apparently a willingness to pay large costs to transform small amounts of cash into larger sums.

Of course, the need to make such transformations is not unique to the poor. And surely some of the phenomena we discuss here may also appear among the middle class. We conjecture, however, that these practices are much more common among the poor in the United States. The well-off have access to a variety of institutions—from store credit for durable purchases to automatic savings deductions—that are intended to facilitate such transformations, which make the well-off less likely to resort to more exotic, and costly, institutions.

No Buffer-Stock Savings Despite High Volatility

One of the fundamental observations of behavioral research is the exceedingly "local" nature of everyday decisions. More-global perspectives, considerations about the long term, are often discounted in favor of issues salient at the moment. Thus, even when long-term decisions are made, they tend to be influenced by minor contextual nuances at the moment of decision that often have little relevance for the long run. Furthermore, long-term forecasts and predictions often fail to take into account the relevance and impact of foreseeable future developments. Along with mental accounting, this typically yields consumption patterns that are overly dependent on current income.

The narrow focusing that emerges has clear implications for planning. Great energy can be spent on decisions of the moment—where to go for dinner or what brand to buy—with relatively little attention allocated to arguably more important decisions that are less immediate, such as how to invest one's retirement savings or whether to save at all. And the failure to plan can be exacerbated when circumstances are highly uncertain and the future less clear, as is often the case in the lives of the poor. With this month's rent proving of great concern, saving for the children's education or for retirement is naturally left until some hopefully better point in the future. The tendency to leave financial planning for a more appropriate moment will be particularly common among those with low incomes, whose finances afford little slack with which to do much planning. An outcome of this highly volatile focus on the moment will be a lack of buffer-stock savings even, or especially, among those people who, in some ways, need it most.

Some Economic Behaviors of the Poor

The Unbanked

A little over 10% of American households are unbanked and have to rely on alternative financial institutions, such as check cashers, to cash in or process their checks (see, e.g., Caskey, 1996; Scholz and Seshadri, 2007). These alternative financial institutions usually charge high fees, and those who use them often have no recourse to formal borrowing instruments. Instead, they resort to high-interest-rate loans, borrow from friends and relatives to make ends meet or to cover emergency spending, or worst, do without access to credit even during tough times.

This costly nonparticipation in banking could be the result of a cost-benefit analysis. If households have little to save, then the benefits of being banked may be outweighed by the financial costs of maintaining an account, such as the minimum-balance fees commonly required by most banks. Alternatively, the choice to remain unbanked could be due to sheer hassle, for example, long travel times, since few banks have branches open in disadvantaged neighborhoods. Low participation may also reflect various cultural factors. Some have attributed to the poor a persistent culture of distrust of financial institutions or have argued that the poor have not internalized a culture of savings and simply prefer living one day at a time, with little planning for the future. What is common to these accounts is a tendency to explain "big" problems, such as millions of unbanked households, through appeal to "big" factors, like a dearth of attractive banking options or a deep mistrust combined with a culture of living from day to day.

In contrast, a behavioral perspective suggests that even in the context of big problems, small factors can sometimes play a decisive role. From a normative perspective, defaults may be largely irrelevant and easily alterable; descriptively speaking, the status quo, bolstered by loss aversion, indecision, procrastination, or a lack of attention, has a force of its own (Samuelson and Zeckhauser, 1988; see also Johnson and Goldstein, this volume). Thus, the mere perception that banks are mostly intended for people of greater wealth may help reinforce the impression that banking is not meant for, and ought not appeal to, those of lesser means. Indeed, decisions that involve being subjected to scrutiny, interviews, requests, and applications, are all likely to have a nontrivial affective component. And those who are most vulnerable are likely to feel the weight of such sentiments even more than the rest. As a number of ethnographic studies suggest (DeParle, 2004; LeBlanc, 2004), the poor often are painfully aware of society's norms and of their own inability to abide by them. A single mother who, without access to child care, needs to present herself at a bank in the company of her small children, may be aware of the fact that, ideally, crying children are not brought into a bank. Or she may worry about her ability to decipher the requisite forms. Along with a severely limited understanding of financial instruments, a poor client may feel reluctance, even shame, and a general sense that she can never be a valued bank customer.

Of course, that perception may not be terribly distant from the truth. There is, after all, a built-in asymmetry in banks' incentives between credit and savings for the poor and the rich. Regarding poorer clients, banks have a greater incentive to promote debt (which can be lucrative, delayed, and compounded) rather than savings (which are bound to be modest), as opposed to the treatment of the wealthy, where debt is likely to be repaid with little penalty and the savings promise to be large and valuable.

In fact, when it comes to bank accounts, the default option is often different for the poor than it is for those who are better off. As mentioned earlier, the simple reliance on direct deposit as a method of payment that gives greater control over one's finances is becoming increasingly popular in the United States, but not among employers of the poor. This misses the opportunity to turn checking accounts into default alternatives for individuals whose de facto default often consists of costly check cashers. Given the power of default options, even among the comfortable, it seems safe to assume that defaults would have at least as substantial an impact on the poor, whose options are inherently inferior, and who may be less informed about alternatives.

From a public-sector perspective, the government could play an important role by further encouraging the automatic transfer of tax (including EITC) refunds to bank accounts. This also provides a way to facilitate the opening of bank accounts. Some evidence from the First Account program in Chicago provides cautious optimism on this front. For many years, the Center for Economic Progress has been providing free tax preparation services for those eligible for the EITC refund. Over the last couple of years, the center has been trying to combine this tax preparation service with the First Account program. Specifically, the center has been singling out individuals who are eligible for a refund but who are without bank accounts. These individuals are then informed that they could get their refund much sooner if they were to open a bank account, to which their refund would be directly deposited. Data obtained from the bank handling the First Account program suggest that those individuals who opened an account in this "quick refund carrot" context were not less likely to still be using their account compared with those individuals (more positively self-selected) who opened an account following a financial education workshop. (See "A Behavioral Perspective on Decisions under Scarcity," below, for further, related findings.)

In light of the discussion above, it is clear that a behavioral view would predict positive effects on saving from the opening of bank accounts. Such accounts should generate a "good" savings default to replace the "bad" money-on-hand situation. In addition, the transfer of cash from, say, checking to savings can trigger a propensity to save more. In fact, bank accounts could be designed specifically to conform to people's "mental accounting" schemes (Thaler, 1999). People

may choose to label one account their "housing account," another their "education account," or yet another their "car account." The labeling of accounts, while nonsensical from the perspective of standard fungibility assumptions, could provide a salient reminder and help with the allocation of specific funds. Such labeling is reminiscent of other, already existing schemes such as education funds, Christmas clubs, and even layaways, and indirect evidence suggests that it may have real consequences. For example, increased child-allowance payments to parents in Sweden were found to have disproportionate effects on the intended recipients' spending on children (discussed in Thaler, 1990).

In contrast with classical analyses, which impute substantial planning and control, numerous studies of middle-class savings suggest that saving works best as a default (see Benartzi, Peleg, and Thaler, this volume; Johnson and Goldstein, this volume). Thus, 401(k)s seem to be effective because the cash is automatically deposited into savings. Yet the poor typically have little recourse to "good" savings defaults. And with good defaults less available without bank accounts, the poor have to revert to alternative, and typically expensive, commitment schemes to try to save toward big purchases. One can view participation in programs such as rent-to-own or layaway schemes as such alternative commitment devices, occasionally leaving the poor in possession of larger amounts than they would be able to save otherwise.

It is fair to note at this juncture that, despite preliminary empirical support, the above proposals would need to be tentatively implemented and seriously evaluated before their full consequences are fully understood. Behavioral outcomes, after all, tend to be multifaceted and complex. Thus, for example, although the appropriate default arrangements may increase savings, it is possible that people with newly automated savings may only come to feel more empowered to take on greater debts, presumably licensed by their new savings. The dynamic and malleable nature of behavior often necessitates pilot testing and evaluation prior to full implementation before the perceived and ultimate impact of new instruments can be fully understood.

Payday Loans

Payday loans are a commonly used financial vehicle amongst lower and middle income households (see Skiba and Tobacman, 2007, and Stegman, 2007, for an analysis). The typical payday loan involves receiving an advance on one's paycheck for a week or two, but this advance comes at a steep price, an effective interest rate that can be as high as 7000+% APR. Such loans are highly contentious from a policy point of view and are often implicitly used to point out the myopia of the poor. We make two basic observations about this widespread institution.

First, as noted above, the highly credit-constrained sometimes find themselves at the edge of poverty. In these circumstances, there may be no myopia in taking out a payday loan. Instead, the local cost-benefit calculus, however painful, may be sound. The lack of cash at crucial times can result in disastrous and mounting consequences—such as having one's telephone service cut off. In these circumstances, even (especially!) the farsighted would take out a loan at high interest rates. The "error" will have happened earlier. It happened through a sequence of actions that left the individual without a buffer stock to deal with shocks. In this view, therefore, there will be circumstances where the puzzle is not why the poor take out payday loans but why they find themselves in situations where they need them.

This perspective poses an interesting challenge to policy makers, who should want borrowers to have access to the loans *at the time of borrowing*. Suppose payday loans are taken by people in severe need, that the need they face is real, and that failure to meet it will have even more severe consequences. Put in this light, payday loans may be a lesser evil compared with policies that use interest-rate caps (or other vehicles) to drive out payday lenders, which can make the poor *worse off*. Unless these are accompanied by policies that solve the lack of a buffer stock among the poor, principled arguments against payday loans are, once again, fundamentally predicated on the expectation that the poor ought to act more "rationally." Instead, such policies can render the poor only more vulnerable to the various shocks they face. Note that a counterargument would be if somehow the lack of availability of payday loans made those who resort to them into better planners. While this is a priori possible, it seems unlikely, and certainly should not be straightforwardly assumed. If despite facing huge consequences, individuals still fail to plan, why would the addition of yet another cost have the desired effect?

To further understand the propensity to resort to such loans, we should ask in what sense payday loans are so very costly. Rather than focus on whether the fees reflect marginal costs or monopoly profits, we should ask, What is the psychologically accurate way to view such costs? Do they actually reflect an individual's net present-value calculation at some exorbitantly high (7000+% APR) rate? Or is the behaviorally most compelling perspective one that suggests more bearable debts? As discussed above, magnitudes are

often evaluated in narrow contexts. People may be willing to travel 30 minutes to save $10 on a $30 purchase but not to save $20 on a $500 purchase. Just as we should not impute a low value of time (less than $20 per hour) from the first behavior or a high value of time (more than $40 per hour) from the second, it may not be right to impute actual discount rates to the intertemporal trade-offs implicit in specific payday loans.

Consider someone who is thinking about paying $20 to get a one-week advance on their $200 paycheck. Such a transaction could be psychologically coded in nominal levels: $20 for a one-week, highly beneficial advance. Viewed in these terms, it may not seem like such a bad transaction. (After all, when the wealthy individual pays $2 to withdraw $100 from an ATM machine out of town, she is really stating a willingness to pay $2—not a general proneness to pay 2% to withdraw her own cash.) Of course, when put into annual rates, the payday loan above implies an APR of over 14000%! The disjunction between the absolute amount and its APR is the result of compounding. But, of course, the individual is not actually making this decision over a year: they typically make this decisions a few times a year, and each for a short period, so the actual compounding is more of a technical than an experienced cost. In short, while the pricing of payday loans may raise economic as well as ethical questions about competition (supply-side issues), psychology can shed light on why individuals would be willing to pay such high rates, without necessarily assuming immense discount rates. Especially for short-horizon loans, computed APRs may not appropriately capture how individuals naturally frame the intertemporal trade-off.

Check Cashing

Like many other services provided to the poor, check cashing is a costly option that provides a service that the well-to-do get for less. In a survey of households living in low- and moderate-income census tracts in Chicago, Los Angeles, and Washington, DC, Berry (2004) found that people often had a fairly accurate understanding of the relative costs of products provided by banks and check cashers. Nonetheless, for many individuals who would be unable to adhere to banks' minimum requirements, costly check-cashing arrangements can prove to be the lower-cost option.

The willingness to engage in costly arrangements may be further facilitated by some of the behavioral proclivities reviewed above. Loss aversion is likely to increase the attractiveness even of fairly costly ways to delay or altogether avoid permanent losses. And the high costs of financial services may be mentally aggregated with perceived gains to which they contribute in the short run, thus leading to an accounting, which at least locally proves more attractive.

While alternatives to costly check cashing often exist, these may be less familiar, less common, and less readily available, especially to low-income participants. A behavioral analysis suggests that it is often not merely the existence of favorable alternatives that will make the greatest difference but that the design of effective channels, combined with directed information, can be critical. For example, in a recent intervention intended to increase elderly Americans' enrollment in Medicare Part D prescription-drug coverage, Kling et al. (2012) documented significantly higher enrollment rates, with an average of at least $230 savings, among participants who were mailed personalized information (regarding their current plan and costs) as compared to a control group who were provided with information regarding the official website where comparable information could be obtained.

In another illustration, credit unions and check cashers in New York have pioneered the use of the point-of-banking machine to facilitate deposits for credit union members at check-cashing stores, providing immediate liquidity of funds and greater convenience for consumers (Stuhldreher and Tescher, 2005). While such arrangements can prove highly beneficial, other partnerships between banks and nonbanks to facilitate payday loans have at times had negative consequences for consumers. Taking the implications of behavioral research seriously, regulators will need to focus on promoting partnerships between banks and nonbanks that provide a more responsive and beneficial range of services to unbanked and underbanked consumers.

A Behavioral Perspective on Decisions under Scarcity

The foregoing analysis made no special allowances for the psychology of the poor. We attributed to people in poverty the same behavioral traits and proclivities expected from those in all other walks of life and explored some of the challenges and exigencies that occur particularly in contexts of poverty. In more recent work, we have begun to explore yet another dimension pertinent to the analysis, which we summarize below.

Scarcity compounded by instability makes everyday life under poverty a constant challenge. Poverty is a domineering context, and the volatility of everyday life affects the poor far more acutely than it affects the well-off. The poor have less economic slack than the

rich. Whereas a rich person can often cut back on (by their own admission) more frivolous spending, a poor person faced with a financially demanding situation is forced to cut back on essential expenses. Living under conditions of poverty poses everyday practical challenges as well as critical demands on mental resources, such as attention, planning, problem solving, and self-control. When these resources are depleted, people make less optimal choices, which, in turn, diminishes their ability to deal with life's frequent challenges, yielding predictable poverty traps. It is the logic and consequences of this cycle of poverty and its psychic challenges that we briefly address next.

Limited Mental Resources

Cognition is a limited resource that exhibits severe and consequential limits of perception and attention. Those challenges on cognitive resources, we suggest, are further impacted by poverty. The cognitive system has a limited capacity, with the processing of each additional stimulus taking up valuable cognitive resources (e.g., Baddeley and Hitch, 1974; Neisser, 1976). The need to deal with problems of scarcity and instability imposes great attentional demands; essentially all financial transactions, no matter how small, must take one's limited budget into careful consideration. In turn, considerably less room is then available for the many other unexpected, threatening, and challenging problems that continuously arise in contexts of instability and lack of slack.

To a large degree, what enters our cognition, or even occupies our attention, is often beyond our control (Wilson, 2004). And in many circumstances, the contexts of the poor provide an especially large array of unwanted and distracting stimuli, including living environments that are loud (Evans, Eckenrode, and Marcynyszyn, 2007), crowded (Evans, Eckenrode, and Marcynyszyn, 2007), and unsafe (Kling, Liebman, and Katz, 2007). Consider, for example, the role that noise plays in impeding children's learning. In one illuminating study, reading test scores of children in classrooms next to passing loud trains were compared to reading test scores of children whose classrooms were in the same school's quiet side. Scores of those on the school's loud side were significantly worse. In response, the school administrators installed noise-absorbing insulation in the noisy classrooms. A follow-up reexamination of student performance showed a reduced difference in test scores (Bronzaft, 1981). The preponderance of evening shift work among low-income workers and the impact of a shift work schedule on sleep has similar effects on life outcomes, with parents exhibiting less tolerance and harsher parenting. (Presser, 2005; Quesnel-Vallée, DeHaney, and Ciampi, 2010).

Cognitive Load

Studies show that performance on critical executive functions, such as cognitive control and working memory, is impeded under greater cognitive load (Lavie, 2000; Lavie et al., 2004). Fockert et al. (2001) show greater interference under conditions of high (vs. low) working-memory load, along with greater related visual cortex activity in the high-load condition. Increased working-memory load has also been shown to reduce the executive control of attention (as gauged via task-switching tasks, inhibitory task performance, and related experimental manipulations; Roberts, Hager, and Heron, 1994).

Cognitive load can influence performance on other important behaviors, as well, such as self-control. Decisions that require self-control are influenced by two competing forces: visceral, present-focused drives, which push people in the direction of succumbing to temptation, and long-term goals that require more effortful, resource-intensive cognitions that help resist temptation (Hinson, Jameson, and Whitney, 2003; Hoch and Loewenstein, 1991; Loewenstein, 1996; Shiv and Fedorikhin, 1999; Sjoberg, 1980; Ward and Mann, 2000). Self-control is weakened to the extent that resources devoted to resisting temptation are exhausted, and research suggests that cognitive load is one such source of weakening. In one set of studies, for example, it was shown that cognitive load disinhibits eating by restrained eaters (Shiv and Fedorikhin, 1999; Ward and Mann, 2000). In another study, load was manipulated by having participants maintain in short-term memory either a two-digit or seven-digit number. Participants were then invited to choose between cake and a fruit salad. As predicted, a significantly greater proportion of those experiencing the greater load opted for the cake (63% vs. 41%), suggesting that cognitive load interferes with people's otherwise regular monitoring of their eating behavior. Other research has documented similar behaviors around time-inconsistency, the apparent alteration of preferences that comes from, among other things, the momentary influence of temptation (Loewenstein, O'Donoghue, and Rabin, 2003).

Load, Stress, and Tunneling: The Importance of Now

As mentioned earlier, a frequently observed behavioral fact is the exceedingly local nature of everyday decisions. More global perspectives and concerns with the long term are often discounted in favor of issues

salient at the moment. This narrow focusing has clear implications for planning. Great deliberation can be given to decisions in the present—which bills to pay first or how to afford a necessary expense—with relatively little attention allocated to important decisions that are less immediate, such as retirement or education planning, or whether to save at all. In fact, everyday focal goals can inhibit the attention given to long-term objectives (Neisser, 1976; Shah, Friedman, and Kruglanski, 2002; Simons and Chabris, 1999). Neglecting priorities as small as adhering to daily medication or the regular payment of bills can easily trigger a chain of events that leads to poorer health or costly financial outcomes. Yet, the very cognitive resources that carry great load under scarcity have been shown to play a pivotal role in helping to resist local temptations in favor of more global interests.

Stress represents an additional factor that can dominate cognition and hamper performance. The relationship between emotional arousal and performance tends to be U-shaped, with greater arousal increasing performance until, when high enough, it starts hampering it (Yerkes and Dodson, 1908). In his cue-utilization theory, Easterbrook (1959) posited that high levels of arousal lead to a restriction of the amount of information to which agents pay attention. The physiological response to stress, referred to as *tunneling*, often helps focus on the object of attention, but focusing on the stressor often comes at the cost of neglecting all else. Tunneling to solve the most pressing immediate problems can create more severe problems in the not-so-distant future. Even when long-term goals are a greater priority, anxiety can cause a person to deploy excessive attention toward threat-related stimuli compared to less anxious individuals, who can distribute their attention more easily (MacLeod and Mathews, 1988; MacLeod, Mathews, and Tata, 1986; Sorg and Whitney, 1992).

Stress is elevated by a variety of stimuli, including noise, bureaucratic tension, and arbitrary discrimination, among others (Glass and Singer, 1972), all of which are disproportionately common in the lives of the poor. Unsurprisingly, the poor experience chronically elevated physiological stress, which has been documented by higher cortisol levels—hormones that increase with stress levels (Sapolsky, 2004). Functioning under a burdened cognitive load, lack of slack, and a variety of persistent stressors, the poor face the risk of ignoring potentially important concerns in favor of those that are pressing at the moment.

Instability and the lack of slack encourage individuals to tunnel—solve today's problems at the expense of tomorrow's. A focus on today, along with a failure to plan, may become only more pronounced when present circumstances are highly challenging and the future uncertain, as is often the case in the lives of the poor. For example, a drop in hours worked this week makes the availability of the rent payment suddenly uncertain. This stressor triggers tunneling: the person does all she can to ensure rent is paid, but this focus forces her to put other priorities—car maintenance, the child's lessons, deposits into a savings account—on the "back burner" until some hopefully easier future time.

Depleted Resources and Self-Control

Self-control, persistence, the delay of gratification, and emotion regulation have all been argued to tap into an exhaustible resource that can be depleted (Muraven and Baumeister, 2000). Consistent with such "ego depletion," persistence on tasks diminishes, and the likelihood of self-control failure increases after a person has exercised self-control. In one illustrative study, for example, hungry participants who have had to refrain from eating cookies (thus depleting self-control) showed less persistence on a puzzle task than those who had not faced the cookie temptation; in another, participants depleted by a difficult, attention-exerting cognitive task were then more susceptible to passive decision-making defaults compared to controls who performed an easier task and were thus less depleted (Baumeister et al., 1998; see also Muraven, Tice, and Baumeister, 1998).

Continuously exerting self-control, resisting temptation, and delaying gratification can be depleting, with deleterious consequences for attention and performance. We might expect the poor, with their lack of slack and a recurring need to show restraint and resist temptation, to be chronically ego depleted. Both poverty and the instability it brings deplete self-control, making saving, planning, and coping with unforeseen events especially challenging. Highly constrained financial resources and a lack of slack require constant budgeting vigilance, resistance, and delayed gratification, all of which drain attention and self-control and further diminish capacity.

Concluding Remarks

The challenges and failures of the poor are a sensitive topic, especially because of the politics involved. In contrast with classical assumptions, the current perspective sheds a somewhat different light. We summarized the accumulated behavioral insights regarding the central role that context plays in shaping human behavior and then considered the variety of ways in

which the context that the poor inhabit provides persistent challenges, along with often insufficient help. This analysis was predicated on the assumption that, contextual differences aside, the psychology of the poor, their behavioral inclinations, and their impulses are not significantly different from those of others.

We then briefly outlined ongoing work exploring the various ways in which the psychology of the poor, precisely because of their particular circumstances, may exhibit other, or at least more extreme, behaviors. Research has documented the profound impact of limited cognitive resources, such as attention and working memory, that are challenged and depleted by the load and demands imposed by trying contexts. Those limitations on resources, we argue, may have particularly pronounced effects in contexts experienced by the poor.

Our perspective here provides a significant departure from the two standard views of the poor. We do not think the behaviors of the poor are inherently different from those of others, nor do we expect their behaviors ultimately to be the same. Instead, we are motivated by research that shows the powerful impact of context on behavior to yield behaviors among the poor that are partly shaped by their unique context. This makes for a fundamentally different perspective: it assumes, among other things, that anyone put in a psychological environment of scarcity would behave similarly. There are several lines of research in support of this hypothesis that involve both the poor and the wealthy (Mullainathan and Shafir, forthcoming).

The present perspective has important implications for policy design and regulation. Low-income decision makers, according to this view, are neither perfectly rational nor particularly incapable, but, rather, are short on "bandwidth"—attention, memory, self-control, and other cognitive resources needed to attend to and act on choices. Policies for the poor, therefore, should regard this population as not merely short on money but also, perhaps especially, on bandwidth. That means doing the same things with regard to bandwidth that we typically would do to address the shortage of money or any scarce resource. In this case, start by paying careful attention to the context in which people function—ranging from financial institutions, benefits programs, and the design of default structures to the availability of child care, transportation, and the complexity of application forms. This perspective, to the extent that it captures important truths, is likely both to enrich and to complicate our views on the role of institutions and of regulation. However, as long as it is founded on a better understanding of decision making and can generate novel and insightful policies, it clearly seems worth the effort.

Notes

Parts of this chapter appeared earlier in Blank and Barr (2009). This chapter was written before Mullainathan joined the federal government, as Director of Research at the Consumer Financial Protection Agency. Nothing here represents, in any way, an official position of the United States government.

1. For more see http://legacy.americanpayroll.org/pdfs/paycard/DDsurv_results0212.pdf

References

Adkins, N. R., and Ozanne, J. L. (2005). The low literate consumer. *Journal of Consumer Research*, *32*, 1, 93–105.

Arkes, H. R., and Blumer, C. (1985). The psychology of sunk cost. *Organizational Behavior and Human Performance*, *35*, 129–140.

Ashraf, N., Karlan, D., and Yin, W. (2006). Tying Odysseus to the mast: Evidence from a commitment savings product in the Philippines. *Quarterly Journal of Economics*, *121*(2), 635–672.

Ausubel, L. M. (1991). The failure of competition in the credit card market. *American Economic Review*, *81*(1), 50–81.

Baddeley, A. D., and Hitch, G. J. (1974). Working memory. *Psychology of Learning and Motivation*, *8*, 47–89.

Barr, M. (2004). Banking the poor. *Yale Journal of Regulation*, *21*, 121–237.

Barr, M. S., Mullainathan, S., and Shafir, E. (2008a). *Behaviorally informed financial services regulation*. New America Foundation White Paper. Washington, DC.

———. (2008b). *An opt-out home mortgage system*. Hamilton Project Discussion Paper 2008-14. Washington, DC: The Brookings Institution.

Basu, K. (2008). *Hyperbolic discounting and the sustainability of rotational savings arrangement*. MPRA Paper No. 20440. Munich Personal RePEc Archive. Retrieved from http://mpra.ub.uni-muenchen.de/20440/1/MPRA_paper_20440.pdf

Baumeister, R. F., Bratslavsky, E., Muraven, M., and Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology*, *74*, 5, 1252–1265.

Benartzi, S., and Thaler, R. H. (2004.) Save More Tomorrow: Using behavioral economics to increase employee saving. *Journal of Political Economy*, *112*(1), 164–187.

Berry, C. (2004). *To bank or not to bank? A survey of low-income households*. Joint Center for Housing Studies Working Paper Series. Cambridge, MA: Joint Center for Housing Studies.

Bertrand, M., Mullainathan, S., and Shafir, E. (2004).

A behavioral economics view of poverty. *American Economic Review*, *94*(2), 419–423.

———. (2006). Behavioral economics and marketing in aid of decision-making among the poor. *Journal of Public Policy and Marketing*, *25*(1), 8–23.

Blalock, G., Just, D. R., and Simon, D. H. (2007). Hitting the jackpot or hitting the skids: Entertainment, poverty, and the demand for state lotteries. *American Journal of Economics and Sociology*, *66*(3), 545–570.

Blank, R. M., and Barr, M. S. (Eds.). (2009). *Insufficient funds: Savings, assets, credit, and banking among low-income households*. New York: Russell Sage Foundation.

Briley, D. A., and Aaker, J. L. (2006a). Bridging the culture chasm: Ensuring that consumers are healthy, wealthy and wise. *Journal of Public Policy and Marketing*, *25*(1), 53–66.

———. (2006b). When does culture matter? Effects of personal knowledge on the correction of culture-based judgments. *Journal of Marketing Research*, *43*, 395–408.

Bronzaft, A. L. (1981). The effect of a noise abatement program on reading ability. *Journal of Environmental Psychology*, *1*(3), 215–222.

Buehler, R., Griffin, D., and Ross, M. (1994). Exploring the "planning fallacy": Why people under-estimate their task completion times. *Journal of Personality and Social Psychology*, *67*(September), 366–381.

Camerer, C., Babcock, L., Loewenstein, G., and Thaler, R. H. (1997). A target income theory of labor supply: Evidence from cab drivers. *Quarterly Journal of Economics*, *112*(2), 407–441.

Caskey, J. P. (1996). *Fringe banking: Check-cashing outlets, pawnshops and the poor*. New York: Russell Sage Foundation.

Croizet, J. C., and Claire, T. (1998). Extending the concept of stereotype threat to social class: The intellectual underperformance of students from low socioeconomic backgrounds. *Personality and Social Psychology Bulletin*, *24*(6), 588–594.

Darley, J. M., and Batson, C. D. (1973). From Jerusalem to Jericho: A study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology*, *27*(1), 100–108.

DeParle, J. (2004). *American dream: Three women, ten kids, and a nation's drive to end welfare*. New York: Viking.

Easterbrook, J. A. (1959). The effect of emotion on cue utilization and the organization of behavior. *Psychological Review*, 66, 183–201.

Edin, K., and Lein, L. (1997). *Making ends meet: How single mothers survive welfare and low-wage work*. New York: Russell Sage.

Evans, G. W., Eckenrode, J., and Marcynyszyn, L. (2007, October). *Poverty and chaos*. Paper presented at the First Bronfenbrenner Conference, Chaos and Children's Development: Levels of Analysis and Mechanisms. Ithaca, NY.

Fockert, J. W., Rees, G., Frith, C. D., and Lavie, N. (2001). The role of working memory in visual selective attention. *Science*, *291*, 1803–1806.

Glass, D. C., and Singer, J. E. (1972). *Urban stress: Experiments in noise and social stressors*. New York: Academic Press.

Higgins, E. T. (2000). Making a good decision: Value from fit. *American Psychologist*, *55*(11), 1217–1230.

Higgins E. T., Idson, L. C., Freitas, A. L., Spiegel, S., and Molden, D. C. (2003). Transfer of value from fit. *Journal of Personality and Social Psychology*, *84*, 1140–1153.

Hinson, J. M., Jameson, T. L., and Whitney, P. (2003). Impulsive decision making and working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 298–306.

Hoch, S. J., and Loewenstein, G. F. (1991), Time-inconsistent preferences and consumer self-control. *Journal of Consumer Research*, *17*, 492–507.

Johnson, E. J., and Goldstein, D. (2003). Do defaults save lives? *Science*, *302*, 1338–1339.

Johnson, E. J., Hershey, J., Meszaros, J., and Kunreuther, H. (1993). Framing, probability distortions, and insurance decisions. *Journal of Risk and Uncertainty*, *7*(1), 35–51.

Kahneman, D. (1994). New challenges to the rationality assumption. *Journal of Institutional and Theoretical Economics*, *150*, 18–36.

Kahneman, D., Knetsch, J. L., and Thaler, R. H. (1990). Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economics*, *98*, 1325–1348.

Kearny, M. S. (2005). State lotteries and consumer behavior. *Journal of Public Economics*, *89*(11–12), 2269–2299.

Kling, J. R., Liebman, J. B., and Katz, L. F. (2007). Experimental analysis of neighborhood effects. *Econometrica*, *75*(1), 83–119.

Kling, J. R., Mullainathan, S., Shafir, E., Vermeulen, L., and Wrobel, M. V. (2012). Comparison friction: Experimental evidence from Medicare drug plans. *Quarterly Journal of Economics*, *127*(1), 199–235.

Knetsch, J. L. (1989). The endowment effect and evidence of nonreversible indifference curves. *American Economic Review*, *79*(5), 1277–1284.

Koehler, D. J., and Poon, C.S.K. (2006). Self-predictions overweight strength of current intentions. *Journal of Experimental Social Psychology*, *42*, 517–524.

Lavie, N. (2000). Selective attention and cognitive control: Dissociating attentional functions through different types of load. In S. Monsell and J. Driver (Eds.), *Attention and performance* (Vol. 18, pp. 175–194). Cambridge, MA: MIT Press.

Lavie, N., Hirst, A., de Fockert, J. W., and Viding, E. (2004). Load theory of selective attention and cognitive control. *Journal of Experimental Psychology: General*, *133*, 339–354.

LeBlanc, A. N. (2004). *Random family: Love, drugs, trouble, and coming of age in the Bronx.* New York: Scribner.

LeBoeuf, R. A., Shafir, E., and Bayuk, J. B. (2010). The conflicting choices of alternating selves. *Organizational Behavior and Human Decision Processes*, *111*(1), 48–61.

Lepper, M. R., Greene, D., and Nisbett, R. W. (1973). Undermining children's intrinsic interest with extrinsic reward: A test of the "overjustification" hypothesis. *Journal of Personality and Social Psychology*, *28*(1), 129–137.

Leventhal, H., Singer, R., and Jones, S. (1965). Effects of fear and specificity of recommendation upon attitudes and behaviour. *Journal of Personality and Social Psychology*, *2*(2), 20–29.

Lewin, K. (1951). *Field theory in social science.* New York: Harper.

Lockley, S. W., Cronin, J. W., Evans, E. E., Cade, B. E., Lee, C. J., Landrigan, C. P., et al. (2004). Effect of reducing interns' weekly work hours on sleep and attentional failures. *New England Journal of Medicine*, *351*, 1829–1837.

Loewenstein, G. F. (1996). Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes*, *65*, 272–292.

Loewenstein, G. F., O'Donoghue, T., and Rabin, M. (2003). Projection bias in predicting future utility. *The Quarterly Journal of Economics*, *118*(4), 1209–1248.

Loewenstein, G., and Prelec, D. (1992). Anomalies in intertemporal choice: Evidence and an interpretation. *Quarterly Journal of Economics*, *107*, 573–597.

Loewenstein, G., and Thaler, R. H. (1989). Intertemporal choice. *Journal of Economic Perspectives*, *3*, 181–193.

MacLeod, C., and Mathews, A. (1988). Anxiety and the allocation of attention to threat. *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 40*(4-A), 653–670.

MacLeod, C., Mathews, A., and Tata, P. (1986). Attentional bias in emotional disorders. *Journal of Abnormal Psychology*, *95*, 15–20.

Madrian, B. C., and Shea, D. F. (2001). The power of suggestion: Inertia in 401(k) participation and savings behavior. *Quarterly Journal of Economics*, *116*(4), 1149–1187.

McCoy, P. A. (2005). A behavioral analysis of predatory lending. *Akron Law Review*, *38*(4), 725–740.

Mendel, D. (2005). *Double jeopardy: Why the poor pay more.* Baltimore: Annie E. Casey Foundation.

Milgram, S. (1974). *Obedience to authority.* New York: Harper and Row.

Mullainathan, S., and Shafir, E. (2009). Savings policy and decision-making in low-income households. In M. Barr and R. Blank (Eds.), *Insufficient funds: Savings, assets, credit and banking among low-income households* (pp. 121–145). New York: Russell Sage Foundation.

Mullainathan, S., and Shafir, E. (forthcoming). *The packing problem: Time, money, and the new science of scarcity.* New York: Henry Holt.

Muraven, M., and Baumeister, R. F. (2000). Self-regulation and depletion of limited resources: Does self-control resemble a muscle? *Psychological Bulletin*, *126*(2), 247–259.

Muraven, M., Tice, D. M., and Baumeister, R. F. (1998). Self-control as a limited resource: Regulatory depletion patterns. *Journal of Personality and Social Psychology*, *74*(3), 774–789.

Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology.* New York: W. H. Freeman.

Presser, H. B. (2005). *Working in a 24/7 economy: Challenges for American families.* New York: Russell Sage.

Quesnel-Vallée, A., DeHaney, S., and Ciampi, A. (2010). Temporary work and depressive symptoms: A propensity score analysis. *Social Science and Medicine*, *70*(12), 1982–1987.

Roberts, R., Hager, L., and Heron, C. (1994). Prefrontal cognitive processes: Working memory and inhibition in the antisaccade task. *Journal of Experimental Psychology*, *123*(4), 374–393.

Ross, L., and Nisbett, R. E. (1991). *The person and the situation: Perspectives of social psychology.* New York: McGraw-Hill.

Rutherford, S. (2001). *The poor and their money.* Oxford University Press.

Samuelson, W., and Zeckhauser, R. J. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, *1*(1), 7–59.

Sapolsky, R. M. (2004). Social status and health in humans and other animals. *Annual Review of Anthropology*, *33*, 393–418.

Scholz, J. K., and Seshadri, A. (2007). *The assets and liabilities held by low-income families.* Paper presented at the conference Access, Assets, and Poverty. National Poverty Center, University of Michigan, Ann Arbor. Retrieved from http://www.nationalpovertycenter.net/news/events/access_assets_agenda/scholz_and_seshadri.pdf

Shafir, E. (2007). Decisions constructed locally: Some fundamental principles of the psychology of decision making. In A. W. Kruglanski and E. T. Higgins (Eds.), *Social psychology: Handbook of basic principles* (pp. 334–352). New York: Guilford Press.

Shafir, E., Diamond, P., and Tversky, A. (1997). Money illusion. *Quarterly Journal of Economics*, *62*(2), 341–374.

Shah, J. Y., Friedman, R., and Kruglanski, A. W. (2002). Forgetting all else: On the antecedents and consequences of goal shielding. *Journal of Personality and Social Psychology*, *83*(6), 1261–1280.

Shih, M., Pitinski, T. L., and Ambady, N. (1999). Shifts in women's quantitative performance in response to implicit sociocultural identification. *Psychological Science*, *10*, 80–90.

Shiv, B., and Fedorikhin, A. (1999). Heart and mind in conflict: The interplay of affect and cognition in consumer decision making. *Journal of Consumer Research*, *26*(3), 278–292.

Simons, D. J., and Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, *28*(9), 1059–1074.

Sjoberg, L. (1980). Volitional problems in carrying through a difficult decision. *Acta Psychologica*, *45*, 123–132.

Skiba, P. M., and Tobacman, J. (2007). Measuring the individual-level effects of access to credit: Evidence from payday loans. *Federal Reserve Bank of Chicago, Proceedings* (May), 280–301.

Sorg, B. A., and Whitney, P. (1992). The effect of trait anxiety and situational stress on working memory capacity. *Journal of Research in Personality*, *26*(3), 235–241.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*(6), 613–629.

Steele, C. M., and Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*, 797–811.

Stegman, M. (2007). Payday lending. *Journal of Economic Perspectives*, *21*, 169–190.

Stuhldreher, A., and Tescher, J. (2005). *Breaking the savings barrier: How the federal government can build an inclusive financial system*. Issue Brief. Washington, DC: New America Foundation.

Thaler, R. H. (1985). Mental accounting and consumer choice. *Marketing Science*, *4*(3), 199–214.

———. (1990). Savings, fungibility, and mental accounts. *Journal of Economic Perspectives*, *4*(1), 193–205.

———. (1992). *The winner's curse: Paradoxes and anomalies of economic life*. New York: The Free Press.

———. (1999). Mental accounting matters. *Journal of Behavioral Decision Making*, *12*(3), 183–206.

Tufano, P. (2009). Consumer finance. *Annual Review of Financial Economics*, *1*, 227–247.

Turner, J. C. (1987). A self-categorization theory. In J. C. Turner, M. A. Hogg, P. J. Oakes, S. D. Reicher, and M. S. Wetherell (Eds.), *Rediscovering the social group: A self-categorization theory* (pp. 42–67). Oxford: Basil Blackwell.

Tversky, A., and Kahneman, D. (1991). Loss aversion in riskless choice: A reference dependent model. *Quarterly Journal of Economics*, *106*(4), 1039–1061.

Van Dort, B. E., and Moos, R. H. (1976). Distance and the utilization of a student health center. *Journal of American College Health Association*, *24*(3), 159–162.

Ward, A., and Mann, T. (2000). Don't mind if I do: Disinhibited eating under cognitive load. *Journal of Personality and Social Psychology*, *78*(4), 753–763.

Willis, L. E. (2006). *Decisionmaking and the limits of disclosure: The problem of predatory lending: Price*. Legal Studies Paper No. 2006-27. Loyola Law School, Los Angeles.

Wilson, T. D. (2004). *Strangers to ourselves: Discovering the adaptive unconscious*. Cambridge, MA: Harvard University Press.

Yerkes R. M., and Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, *18*, 459–482.

CHAPTER 17

Psychological Levers of Behavior Change

DALE T. MILLER

DEBORAH A. PRENTICE

How can we change people's behavior, change it deliberately, with a specific goal in mind? This is a central question for social scientists and policy makers who seek to ameliorate societal problems, for, of course, human behavior is at the heart of many of these problems. If people would just stop driving SUVs, keeping their houses so warm, tossing recyclables into the trash, and leaving lights and appliances on, carbon emissions would fall substantially. If they would watch what they eat, do their 30 minutes of exercise each day, wear sunscreen, buckle up, and drink in moderation, they would have a better quality of life and, at the same time, produce less pressure on the health-care system. If they would sign up for and contribute to 401(k) programs, take advantage of the tax breaks offered by health-care and dependent-care expense accounts, and keep their credit card debt down, their money would go further. Indeed, if people would just act in ways that align with their own long-term interests and stated convictions, many collective problems would be a lot less pressing. How can policy makers help people to act in ways that further what those people want, for themselves and for society?

Framed in this way, behavior change is a psychological project, one that focuses specifically on the motivations that underlie behavior. The problem in these cases is not what people want; in fact, virtually everybody wants to clean up the environment, live longer and healthier lives, be good to their fellow citizens, get as much as they can for their money, and help those in need. Nor is the problem the know-how; in most of these cases, people know of at least something they could do to further the cause. The problem is motivation: people fall short. There is a gap between what they know they could or should do and what they actually do. Thus, the behavior change that is needed, at least at the individual level, involves getting people to stop doing things they know they should not be doing and start doing things they know they should be doing. Policies need to help people close the gap between aspiration and action.

In this chapter, we consider strategies for producing this kind of behavior change on a large scale, analyzing various intervention efforts in light of their capacity to move people toward desirable behavior and away from undesirable behavior. Invariably, these efforts aim to restructure material and psychological incentives, often in very clever ways. However, as we will argue, their success depends ultimately on two rather subtle and often overlooked factors: (1) the motivational structure of the status quo before the intervention is undertaken, and (2) the complex and nonadditive relationship between material and psychological incentives. We will begin with a brief overview of how psychologists have traditionally approached the task of changing behavior, which provides the conceptual grounding for our subsequent analysis.

Psychological Approaches to Behavior Change

Psychologists have a long history of attempts at changing behavior. Interest in the question of how to design behavioral interventions emerged with the field of social psychology during World War II and has remained strong ever since (see, e.g., Walton and Dweck, 2009). By intervention, we refer to a deliberate attempt by an agent to change the behavior of a group of individuals, a group that can range in size from a couple of people to an entire nation. What is distinctive about the way psychologists approach this task is that they assume that some psychological construct or process controls the behavior in question, and that engaging or modifying that process will, in turn, modify the behavior. Examples of psychological constructs that influence people's behavior

are attitudes, expectations, self-esteem, self-concepts, goals, and identities. Examples of psychological interventions are (1) efforts to reduce intergroup animosity and promote intergroup tolerance by changing attitudes about other groups, (2) efforts to improve individuals' educational performance by boosting self-esteem, (3) efforts to reduce unhealthy life styles by modifying beliefs about the riskiness of certain behaviors, and (4) efforts to promote environmental conservation by fostering a "green" identity.

Of course, psychologists are not the only ones in the business of trying to change behavior. Social scientists of all stripes seek ways of moving people in the direction of social goods, often quite successfully. What is distinctive about the way psychologists approach this task is their sensitivity to the psychological constructs and processes that produce behavior change. For many policy makers—those who seek behavior change through modifications to law, markets, human, or social capital, for example—this level of mechanistic detail remains unexamined. For psychologists, these details are the key to producing a successful campaign. Indeed, the failure to attend adequately to how interventions work at the psychological level has been responsible for many an unsuccessful campaign.

Psychological approaches to behavior change fall into two general categories, each grounded in a distinct intellectual tradition. One is the attitude-change approach, which seeks to change people's attitudes and beliefs in order to change their behavior. This approach is most closely identified with Carl Hovland and the Yale Communication and Attitude Change Program of the 1950s (Hovland, Janis, and Kelley, 1953; see McGuire, 1996, for the history of this program). It has its intellectual origins in the behaviorist movement, which dominated the field of psychology through the middle of the twentieth century, and in the cognitive revolution that supplanted it. For Hovland and the many who have followed in his footsteps, changing behavior is fundamentally about changing minds; it is about controlling the information to which people are exposed and how they process that information. The Hovland tradition has been highly generative and remains prominent, especially in the fields of communications, marketing, and consumer behavior.

A second approach to behavior change, and the one most relevant to the present analysis, focuses not on attitudes and beliefs, but rather on motives. It is most closely identified with Kurt Lewin, a German émigré whose ideas were influenced by various European schools of thought, most notably the Gestalt tradition and the Frankfurt School. Lewin himself was involved in many interventions during World War II, including efforts to strengthen the morale of the fighting troops and to reorient the consumption of Americans away from foods that were in short supply. After the war, he became involved in interventions designed to reduce prejudice, and it was out of this work that the idea of sensitivity training emerged, resulting in the establishment of the National Training Laboratory, or the T-group movement (Ross and Nisbett, 1991). Lewin recognized that behavior is a function not just of attitudes and beliefs but also of motivational dynamics of which individual actors are often only dimly aware. Interventions that target these motivational dynamics can be very effective at getting people to do what they are already convinced that they should and want to do.

Approach and Avoidance Motivation

The motivational distinction on which we draw in this chapter grows directly out of Lewin's (1951) theoretical framework. For Lewin, behavior occurs in what he called a force field (or, for the individual, a life space), where multiple pressures operate simultaneously on the individual. Some of these pressures push the person in the direction of an outcome or goal—these are sources of *approach motivation*; other pressures push the person away from that outcome or goal—these are sources of *avoidance motivation*. Behavior represents the current equilibrium of the system, the point at which approach and avoidance motivations balance each other out in a given moment. All behaviors, even the most simple and routinized, occur in a tension system, and changing a behavior (i.e., changing the equilibrium of the system) requires a change in the force of the different tensions.

The concepts of approach motivation and avoidance motivation may seem like two heads of the same coin, but they are not. To illustrate the distinction between them, consider a (quaintly dated) quotation from Ernest Dichter, a pioneer in advertising research:

> We are now confronted with the problem of permitting the average American to feel moral . . . even when taking two vacations a year and buying a second and third car (cited in Coontz, 1992, p. 171).

What Dichter is saying here is that Madison Avenue has done what it can with the lever of approach motivation and that they now have to focus on reducing the avoidance motivation—that is, guilt—that is keeping people's approach motivation from expressing itself in behavior. In essence, the task is not to motivate consumption; it is to license consumption. Of course, one could argue that if advertisers were able to increase the desire for their products sufficiently, people

would get over their inhibitions. Although that may very well be true, it still might be more economical to invest in reducing avoidance motivation than in increasing approach motivation.

The advertising example illustrates several important points. First, intervention efforts should begin with an analysis of the current balance and sources of approach and avoidance motivation that are affecting the target behavior. Such analysis will reveal the motivational basis for the status quo and may also indicate which intervention strategies are most likely to be productive. On this latter point, Dichter's remarks suggest that interventions may often be more effective and efficient if they seek to move motivation levels away from the extremes. For example, if the goal is to induce a behavior and approach motivation is already high, one might do well to reduce the avoidance motivation that is inhibiting the behavior. However, if the approach motivation is low—if the product or service one is marketing is new or unknown to the consumer, for example—then raising the approach motivation may be a more effective strategy. Likewise, if the goal is to curb a behavior and the avoidance motivation is already high, one might do well to reduce the approach motivation; however, if the avoidance motivation is low, then increasing it may be a more effective strategy. More generally, interventions can utilize either a motivating psychology, by targeting the forces that impel behavior, or a licensing psychology, by targeting the forces that inhibit behavior. On rare occasions, a well-designed and inspired intervention may be able to do both, as we will illustrate.

Levers of Behavior Change

Thus far, we have outlined two questions one must ask when designing an intervention. First, what is the behavioral target of the intervention? Second, what is the psychological target of the intervention? Will it seek to motivate behavior change or to license the existing behavior? Now we would like to introduce a third question: What will be the lever of change? We propose that designers of interventions have two broad categories of levers at their disposal: taxes and subsidies. We use the language of economics here but intend the concepts more broadly—to include psychological, as well as material, taxes and subsidies (see Sunstein, 1996a, 2008). Common examples of psychological taxes are losses of self-respect, public respect, identity, and self-esteem; common examples of psychological subsidies are boosts to this same set of psychic goods.

Interventions designed to change people's behavior can add or remove taxes, and they can also add or remove subsidies. Which strategy or strategies one should use depends, we argue, on the motivational tension system underlying the status quo. An analysis of several well-known interventions serves to illustrate this point.

Don't Mess with Texas

In 1986, the Department of Transportation in the state of Texas, with the help of an Austin-based advertising firm, initiated the Don't Mess with Texas campaign to reduce roadside littering. In this campaign, the phrase "Don't Mess with Texas" is prominently displayed on road signs and in television and print ads and is also used in radio announcements, which are often made by Texas celebrities. The campaign has been highly successful, reducing roadside littering by 72% between 1986 and 1990 (McClure and Spence, 2006). It remains in full swing, with a website (http://dontmesswithtexas.org/) that provides people with opportunities to participate in clean-ups, join partnerships, sponsoring activities, and reporting a litterer. The pitch it gives for undertaking these activities: "It's what real Texans do."

This campaign provides a straightforward example of a successful effort to modify people's motivation to litter. Its goal is to make people more uncomfortable littering, not to make them more comfortable not littering. In other words, the campaign targets a motivating dynamic, not a licensing one. The lever it uses is a tax, a psychological tax on littering conferred by the slogan, "Don't Mess with Texas." Note that this slogan is effective at imposing a tax on littering only because being a good Texan features prominently in the self-identity of the target group. This strategy would not be nearly as effective in regions where state identity is less strong. Consider, for example, the prospects of a Don't Mess with New Jersey campaign. Having lived in New Jersey for many years, we can assert with a considerable degree of confidence that such a campaign would not rival the success of Don't Mess with Texas. The more general point is that one needs to have a good understanding of the target group to design an effective intervention to change its behavior.

Friends Don't Let Friends Drive Drunk

In 1983, the National Highway Traffic Safety Administration, in partnership with the Ad Council, launched a campaign to reduce drunk driving. Their most successful tagline, introduced in 1990, was "Friends Don't Let Friends Dive Drunk." In the first year that this tagline was included at the end of public-service announcements, alcohol-related fatalities fell

nationally by 10%. The slogan went on to become the most highly recognized antidrunk-driving message in U.S. history.

How does this intervention influence behavior? There are at least two possible mechanisms at work. The first is that it capitalizes on a motivating psychology: it seeks to reduce the incidence of drunk driving by increasing people's motivation to stand in the way of would-be drunk drivers. More specifically, it tries to make people feel uncomfortable letting others drive drunk and does so by imposing a psychological tax on this behavior. It says that you are not a good friend if you let a friend drive drunk.

But consider a second possibility. Perhaps this intervention capitalizes not on a motivating psychology but on a licensing psychology: perhaps it seeks to reduce the incidence of drunk driving by reducing people's inhibitions against standing in the way. This mechanism is quite plausible, because people are notoriously reluctant to stand in the way, to challenge another person's face, in Goffman's (1959) terms. Such behavior carries with it significant psychological, and sometimes social, costs. Under this interpretation, the friends-don't-let-friends-drive-drunk intervention seeks to make people feel more comfortable intervening and does so by removing the psychological tax on this behavior.

The beauty of this particular intervention, and perhaps one of the reasons for its success, is that it can work both ways simultaneously. What the intervention does effectively is link the desired behavior with an identity everyone cares about: that of being a good friend. For people who currently fail to stop their friends from driving drunk because they are unmotivated—that is, for those who do not care if others drive drunk—the intervention provides the motivation to act. For people who currently fail to stop their friends from driving drunk because they are inhibited, the intervention provides the license to act.

Social-Norms Marketing

Another behavioral domain in which interventions have sought to leverage social motives is the domain of excessive drinking on college campuses. Here, the state of the art is the so-called social-norms marketing approach, which involves giving students accurate data on the prevalence of, and sometimes attitudes toward, heavy drinking on campus. The goal of these programs is to correct students' mistaken impression that everybody drinks to excess and therefore that drinking in moderation will be frowned upon. Thus, the programs seek not to reduce the actual psychological tax on moderation but to correct the perception of the tax that actually exists on moderation. In short, these campaigns aim to show students that if they drink in moderation, they are not incurring the tax they think they are.

The social-norms marketing approach has been used extensively, with mixed results (see Prentice, 2008, for a review). The approach is clearly designed to reduce students' inhibitions about deviating from what they mistakenly see as a campus norm; it is based on the assumption that the reason students engage in excessive drinking practices is that they do not feel comfortable saying no. This is clearly the case for some of the students some of the time. However, for many students on many campuses, drinking is more like littering—they feel unconflicted about it. The task, given this motivational landscape, is to make students feel more uncomfortable drinking heavily, not to make them more comfortable not drinking heavily.

Nonadditive Effects of Psychological and Economic Incentives

We have argued that psychological taxes can be just as costly, and therefore just as effective at deterring behavior, as economic taxes are. We now turn to a consideration of the relation between psychological and economic incentives. In some cases, these two types of levers combine additively, as in the Don't Mess with Texas campaign, where existing fines on littering and the appeal to Texas identity both pull in the same direction. Psychological and economic subsidies can also combine additively, as in cases in which people get both economic and psychological benefits for doing good—so-called double-bottom-line situations. In many cases, however, the relation is more complicated. Of particular concern here are the various ways in which economic taxes and subsidies can have perverse behavioral effects because of their effects on psychological taxes and subsidies.

Counterintuitive Effects of Economic Taxes

Sometimes the imposition of an economic tax on the desired behavior can actually increase its supply by removing a psychological tax on it. Consider, for example, the introduction of cake mixes in the 1950s. Testing indicated that homemakers liked the look and taste of cakes made from mixes and said they would be happy to serve these to their families. When the manufacturer released this product, however, sales were underwhelming. People simply did not buy them. Further investigation revealed the problem: homemakers felt uncomfortable using cake mixes because they did not feel as if they were really baking. They wanted baking made easier, but they also wanted

what they did still to be considered baking. So what the manufacturers did was to incorporate a few more steps in the process. Now the mixes required users to add eggs, oil, and milk and to beat the batter for several minutes. This change provided a big boost to sales. Why did it work? The change did not improve the appearance or taste of the cakes, and indeed, it increased the amount of effort required to use the product. But ironically, increasing effort was the key to its success. The problem here was not to motivate behavior; it was to license it. Homemakers wanted to use cake mixes; the motivation was there. However, they did not feel licensed to use the mixes until a material tax (in this case an effort tax) was added that enabled them to use the mixes without undermining their identities as homemakers (Etzioni, 1990, p. 69).

Sometimes the imposition of an economic tax on the undesired behavior can decrease its supply by removing a psychological tax on a more desirable alternative behavior. Here, we are indebted to legal scholars Lawrence Lessig (1995, 1996, 1998) and Cass Sunstein (1996a, 1996b), who have provided many examples of how the imposition of legal constraints can reduce psychological taxes and thereby license desirable behavior. One particularly compelling example is the case of dueling in the South (Lessig, 1995). Eradicating dueling in the southern United States proved extremely difficult to do. Dueling was a means of resolving disputes or matters of honor for Southern men of a certain social standing. Even though participating in duels was often rational for the duelists (as a means of attaining higher status), it came with a high collective cost, or at least the cost was perceived as high by the states that struggled to ban the practice. High legal taxes on dueling, which sometimes included death, had little deterrent value. One problem was the difficulty of enforcing the law. Another was the steep psychological tax incurred by refusing the challenge to a duel. Anyone refusing to duel was branded a coward and no gentlemen; no material tax on conviction for dueling could compete with that. Indeed, Southern males of a certain class simply could not say, "I would love to duel you but it is against the law," because the code of honor was meant to be above laws made by commoners.

The solution, or at least a more effective solution, according to Lessig (1995), was to make disqualification from holding public office one of the penalties for a dueling conviction. Why was this effective? One of the obligations of the Southern gentlemen was a willingness to hold public office. Conflicted duelers were unable to fulfill that obligation. Thus, refusing to duel could now be viewed as consistent with, or at least not entirely inconsistent with, the code of honor. The psychological tax on refusal was lower. In short, the law functioned not so much to constrain Southern males from dueling as to license them not to duel.

Note that the effectiveness of this strategy could be expected to grow in parallel with the desire among members of the targeted group to extract themselves from the pressure to duel. With Southern gentlemen increasingly seeking a way out, this particular penalty, and its licensing properties, became increasingly effective. Indeed, one reason different interventions or remedies are effective later when they were not effective earlier is that the motivational underpinnings of behavior have changed. At one time, people might be comfortable doing what they do, but at a later time, they might simply be acting as they always have, even though their feelings have changed. This dynamic, known as a conservative lag, helps to account for why public practices often persist long after they have lost private support (Fields and Schuman, 1976).

Lessig (1995) gave many other examples of how the introduction of laws was effective because those laws changed the meaning of behavior, thereby licensing people to do what they wanted to do but had been inhibited from doing. One interesting implication of this analysis is that, under some circumstances, people might advocate for laws that would constrain them. That is, they might seek a material tax so as to liberate themselves from a psychological tax. This dynamic can help to explain the somewhat puzzling role that Southern businesspeople played in the 1960s civil rights legislation. According to Lessig (1995), in the hearings surrounding this legislation, many business, restaurant, and hotel owners in the South, all of them white, testified in support of legislation that would make it illegal to do what they currently did—namely, not serve African Americans. Why did they want to be forced to do what they would not do voluntarily? The answer, we maintain, can be found in the psychological tax they incurred by voluntarily serving African Americans. Voluntarily serving African Americans made them vulnerable to accusations that they were too greedy or perhaps too sympathetic to blacks. This psychological tax deterred them from doing something that was in their economic interest. Antidiscrimination laws, then, licensed them to do what they wanted to do: to serve as many customers as they could. In effect, they were asking for one kind of tax to free them from a more punitive one.

In summary, a legal intervention's effectiveness depends on its effects on the psychological levers controlling behavior. Most commonly, or at least most intuitively, the legal imposition of a material tax generates additional psychological taxes, including the stigma of being a scofflaw or a criminal. However, sometimes the material tax reduces psychological taxes by legitimating a desired but inhibited behavior.

Sometimes the imposition of an economic tax on the undesired behavior can increase its supply by removing a psychological tax on it. An excellent example of this dynamic is provided by an intervention study conducted in ten private day-care centers in Israel (Gneezy and Rustichini, 2000). The targeted group was not the children but their parents, many of whom were routinely tardy to pick up their children. The intervention strategy was one of traditional deterrence: the investigators imposed a tax on arriving late. At six of the day-care centers, they posted the following announcement:

> As you all know, the official closing time of the day-care center is 4 PM every day. Since some parents have been coming late, we . . . have decided to impose a fine on parents who come late to pick up their children. As of next Sunday a fine of NIS 10 [about $4] will be charged every time a child is collected after 4:10 PM. This fine will be calculated monthly, and it is to be paid together with the regular monthly payment.

The remaining four day-care centers served as a control group. The fine was imposed in the fifth week of the twenty-week observation period and was removed in the seventeenth week.

The results were striking. The number of late pickups *increased* significantly with the imposition of a fine and remained at the higher level even after the fine was removed. What accounts for this perverse effect? One interpretation is that the introduction of an economic tax reduced the psychological tax that parents had been paying when they picked up their children late, a tax that was sufficiently high that most parents did not want to pay it. That is, before the fine, tardy parents were inflicting costs on the other parents and on the day-care-center staff—they were being free riders—and this behavior cost them psychologically through the damage it did to their private reputation, their public reputation, or both. Once a fine was introduced, latecomers were no longer free riding; they were simply paying a material price for their lateness.

This intervention prompts two additional observations. First, note that there surely would have been some price at which the tax would have had a deterrent effect, though not necessarily a price that would have been feasible to implement. Second, it is quite possible that this intervention did not have the same effect on all members of the targeted group: it may have constrained those who had previously come late but liberated a new (and larger) set of latecomers. That is, the original free riders may have been stimulated by the fine to start showing up on time, but a larger group of parents, who had previously been constrained by the moral implications of latecoming, may have been freed up by the fine. The more general point here is that to the extent that the motives that drive status-quo behavior vary across members of a targeted group, so too will the effects of new incentives.

To summarize this discussion of taxation, we can say the following: imposing a material tax can have the effect of removing a psychological tax. This tax will increase the desired behavior, but only if the existing psychological tax was suppressing the desired behavior (as was the case for dueling in the South). If the existing psychological tax was actually encouraging the desired behavior, introducing a material tax can lead to a reduction in the desired behavior (as was the case for tardy day-care pickups).

Counterintuitive Effects of Economic Subsidies

Economic and psychological subsidies operate very similarly to economic and psychological taxes. Often the two combine additively, as, for example, when one receives both economic and psychological gains for contributing to the public good. Tax deductions for charitable giving would be one example; the case of social investing would be another. However, they do not always combine additively, as the following examples illustrate.

Sometimes the introduction of an economic subsidy for the desired behavior can decrease its supply by removing an existing psychological subsidy for it. Consider the case of blood donation. In the early 1970s, when Britain was considering trying to increase its blood supply by moving from a strictly charitable system to a compensated system, Richard Titmuss, a social policy authority, wrote a book arguing that this move would have adverse effects on both the quantity and quality of the blood supply (Titmuss, 1971). In this book, he argued that compensating blood donation would make it less attractive to people because the compensation would deprive them of the gratification they received from what heretofore had been an act of civic virtue. In other words, he argued that the offer of an economic subsidy would reduce the psychological subsidy donors received. Moreover, he maintained that the magnitude of the economic subsidy would almost certainly not be high enough to compensate for the loss of the psychological subsidy. These claims generated a great deal of controversy, in part because Titmuss himself offered little data in support of them. However, recent theory and evidence have borne Titmuss out, at least to some extent (Mellstrom and Johanneson, 2008). For blood donation, as well as for any other behaviors linked to civic virtue, the psychological benefit people receive from the action is undermined if they begin to receive economic benefit from the action as well.

Another example of this paradoxical effect is provided by not-in-my-backyard, or NIMBY, projects,

which include transportation improvements (rail lines and airports), power plants, wastewater treatment plants, landfills, prisons, half-way houses, and homeless shelters. These so-called noxious facilities present a collective action problem, in that they provide benefits for the collective but come at a cost to those neighborhoods that host them. The consequence, of course, is that everyone resists them. The traditional solution to this problem is to tax all those who benefit from the facility and use that money to subsidize those who accept the facility in their neighborhood. However, this approach has not been very successful in getting communities to accept NIMBY projects, and in many cases it has produced lower acceptance rates than no compensation (Frey and Jegen, 2001; Frey and Oberholzer-Gee, 1997).

Why are subsidies ineffective in this case? One problem is the signal that is conveyed by offering compensation—the signal that no one wants this facility and thus it must be highly undesirable. The act of accepting the compensation also signals that one can be bought or bribed: it brands one the kind of person who would put his or her children at risk for money. The economic subsidy, thus, reduces any psychological subsidy that derives from accepting a NIMBY project out of civic-mindedness and could even introduce a psychological tax on acceptance (i.e., the stigma of being bribable). Of course, a large-enough subsidy might compensate for these psychological taxes. However, an alternative form of subsidy appears to work more efficiently: NIMBY projects fare better when communities are offered compensation in a restricted form, for example, in the form of new parks or school improvements (Kahan, 2003). Even though from a rational-actor perspective the offer of compensation in a restricted (nonfungible) form would seem less attractive, in this case nonfungibility increases attractiveness because it reduces the psychological tax on accepting NIMBY projects. It licenses communities to accept these projects (and the associated compensation) without being seen as, on the one hand, a sucker, or, on the other hand, a sellout.

The general question of the relationship between economic subsidies and behavior change has received increasing scrutiny with the growing popularity of conditional cash transfer (CCT) programs (Fiszbein and Schady, 2009). These programs are typically employed in communities where there is deemed to be an underinvestment in human capital, usually education and preventive health care. Most commonly, these programs are designed to increase parents' promotion of their children's education and health by providing the parents with economic incentives conditional on specific outcomes, such as school attendance and visits to public-health sites. The results of these programs have been mixed (Lomeli, 2008). For example, CCT programs have had more consistent positive effects on school attendance than on educational outcomes; similarly, they have had more consistent positive effects on frequency of visits to health clinics than on health outcomes. The question arises, When does seeking out opportunities to improve one's education or one's health lead to such improvements and when does it not?

We cannot claim to have a complete answer to the question of when and where conditional cash transfers will be effective, but the analysis we have presented suggests some guidelines. One key issue, for example, is whether the current level of the problematic action or inaction is best represented as due to a deficit of approach motivation or a surplus of avoidance motivation. For example, are students not attending school because they are not interested in attending school, or are they interested but inhibited by, let us say, an oppositional ideology prevalent in their peer group? If the latter were the case, even a small amount of economic compensation to parents or students might prove effective, because it would reduce the psychological tax imposed by the oppositional ideology—that is, it would license students to both go to school and do well there. If, on the other hand, students were interested in attending school but were inhibited by structural barriers, then compensating school attendance might even reduce it by reducing the intrinsic interest of those students who managed to go despite the structural barriers (see Deci, Koestner, and Ryan, 1999). Again, what all of these examples illustrate is the importance of understanding the field forces operating on the behavior prior to the introduction of subsidies.

Sometimes subsidizing a certain level of consumption can reduce the likelihood of higher consumption levels by imposing a psychological tax on it. Personal use of office supplies—so-called pilferage—imposes a $40-billion-a-year cost on American businesses (http://www.missouribusiness.net/sbtdc/docs/problem_employee_theft.asp). Confronting this problem has proven difficult because most interventions offend employees and prove counterproductive.

One solution that has been shown to be effective is to give employees an allotment (a subsidy) of supplies that constitutes some percentage of what the per capita loss typically is (Kahan, 2003). For example, if the average employee typically takes ten pens per year from the work place, the employer might give each employee an upfront annual allotment of six pens. With this allotment, the total annual loss of pens is likely to be less than the previous average. Why? The answer lies in the effect of the allotment on the meaning of taking the seventh and all subsequent pens. When they have been given six pens, employees no longer feel as entitled to that seventh pen as they

previously did. The allotment introduces a psychological tax on any additional pens taken.

Calibrating Taxes and Subsidies

In many of the foregoing examples, economic taxes and subsidies failed, at least in part, because they were set too low. This is often the case. Material incentives are subject to a host of symbolic, practical, and political constraints that keep their values low. Psychological incentives are typically subject to less scrutiny, although their values are limited by practical, and sometimes ethical, considerations as well. Given this reality, we would like to underscore a point we have made several times in passing: releasing or unfreezing behavior often requires less of an intervention than motivating it does. For example, consider the effect that the first person leaving a social event has on the number and rate of other people leaving. If people are enjoying themselves and are reluctant to leave, the effect of a single person leaving is likely to be minimal and highly dependent on the status and visibility of that person. On the other hand, if people are miserable and desperate to leave, the defection of a single person, no matter how unassuming and low in status he or she is, might prompt a mass exodus. In the latter case, the model functions not to create the impulse to leave but merely to license one that already exists.

This observation suggests a modification to a point we made earlier. In our discussion of the success of the Don't Mess with Texas antilittering campaign, we speculated that a Don't Mess with New Jersey campaign would not be nearly as effective, because most residents of New Jersey do not identify strongly with their state. This claim would be true if the function of the campaign was to motivate people not to litter. However, if the situation on the ground was that residents really did not want to litter but felt pressured to do so by social norms (a situation that might actually obtain in some areas of New Jersey), then a Don't Mess with New Jersey appeal might be sufficient to release them from their inhibitions. More generally, when the function of laws is to remove psychological taxes and license people to do what they want to do anyway, the laws do not need to be very punitive or reliably imposed to be effective.

Summary

In summary, we have offered one general and two specific arguments. The general argument is that efforts to change behavior must begin with a careful analysis of the external and internal circumstances bearing on the status quo. This analysis often, if not always, depends critically on the collection of data. A good illustration of this point is provided by what is perhaps the one, uncontested, positive outcome of the U.S. military's 2007 so-called surge in Iraq: the turning of many Sunni Al Qaeda collaborators into U.S. collaborators. This trend began after General Petraeus conducted a survey of Sunni detainees and found that the most common reason they reported for collaborating with Al Qaeda was to make money to buy luxury goods (Ricks, 2009). This finding led Petraeus to intervene by successfully outbidding Al Qaeda, something the United States could easily do. Of course, this strategy is particular to the situation in Iraq; it is not necessarily going to work in other insurgency situations. But that is precisely our point: interventions must always be particular to the situation. Moreover, it is clear that the outbidding strategy would never have been undertaken without a close analysis of the situation on the ground, an analysis that was strikingly absent during the first four-plus years of the war.

Our specific arguments concern motivational dynamics and the levers that can be used to change them. We have argued that the critical first step in the design of any intervention strategy is to assess the sources and strengths of the approach motivation and avoidance motivation underlying the status quo. An analysis of the motivational dynamics currently operating will dictate whether the challenge is to motivate or to license behavior, which, in turn, will point in the direction of either taxes or subsidies. But, of course, taxes and subsidies also require a thorough analysis. Our final argument is that, in deciding on how to utilize these levers, one must recognize that economic and psychological incentives often combine in complex ways. Our analysis of these complexities has taken full advantage of the benefits of hindsight; a pressing agenda item, for researchers and would-be interveners alike, is to deepen their understanding of the psychology of incentives so that they can predict these complexities.

References

Coontz, S. (1992). *The way we never were: Families and the nostalgia trap.* New York: Basic Books.

Deci, E. L., Koestner, R., and Ryan R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, *125*(6), 627–668.

Etzioni, A. (1990). *The moral dimension: Toward a new economics.* New York: Free Press.

Fields, J. M., and Schuman, H. (1976). Public beliefs about the public. *Public Opinion Quarterly*, *40*, 427–448.

Fiszbein, A., and Schady, N. (2009). *Conditional cash transfers: Reducing present and future poverty*. Washington, DC: The World Bank.

Frey, B. S., and Jegen, R. (2001). Motivation crowding theory. *Journal of Economic Surveys*, 15(5), 589–611.

Frey, B. S., and Oberholzer-Gee, F. (1997). The cost of price incentives: An empirical analysis of motivation crowding-out. *American Economic Review*, *87*(4), 746–755.

Gneezy, U., and Rustichini, A. (2000). A fine is a price. *Journal of Legal Studies*, *22*, 1–17.

Goffman, E. (1959). *The presentation of self in everyday life*. Garden City, NY: Doubleday-Anchor.

Hovland, C. I., Janis, I. L., and Kelley, H. H. (1953). *Communication and persuasion*. New Haven, CT: Yale.

Kahan, D. (2003). The logic of reciprocity: Trust, collective action, and the law. *Michigan Law Review*, *102*, 71–103.

Lessig, L. (1995). The regulation of social meaning. *University of Chicago Law Review*, *62*(3), 943–1045.

———. (1996). Social meaning and social norms. *University of Pennsylvania Law Review*, *144*(5), 2181–2189.

———. (1998). The New Chicago School. *The Journal of Legal Studies*, *27*(2), 661–691.

Lewin, K. (1951). *Field theory in social science* (edited by D. Cartwright). New York: Harper.

Lomeli, E. V. (2008). Conditional cash transfers as social policy in Latin America: An assessment of their contributions and limitations. *Annual Review of Sociology*, *34*, 475–499. Palo Alto, CA: Annual Reviews.

McClure, T., and Spence, R. (2006). *Don't Mess with Texas: The story behind the legend*. Idea City Press.

McGuire, W. J. (1996). The Yale Communication and Attitude-Change program in the 1950s. In E. E. Dennis and E. Wartella (Eds.), *American communication research: The remembered history. LEA's communication series* (pp. 39–59). Hillsdale, NJ: Erlbaum.

Mellstrom, C., and Johannesson, M. (2008). Crowding out in blood donation: Was Titmuss right? *Journal of European Economic Association*, *4*, 845–863.

Prentice, D. A. (2008). Mobilizing and weakening peer influence as mechanisms for changing behavior: Implications for alcohol intervention programs. In M. J. Prinstein and K. A. Dodge (Eds.), *Understanding peer influence in children and adolescents* (pp. 161–180). New York: Guilford.

Ricks, T. (2009) *Gamble*. New York: Penguin.

Ross, L., and Nisbett, R. E. (1991). *The person and the situation: Perspectives of social psychology*. New York: McGraw Hill.

Sunstein, C. R. (1996a). Should government change social norms? *AEI Bradley Lecture Series*. Washington, DC: American Enterprise Institute. Retrieved from http://aei.org/speech/society-and-culture/poverty/should-government-change-social-norms/

———. (1996b). Social norms and social roles. *Columbia Law Review*, *96*, 903–968.

———. (2008). Adolescent risk-taking and social meaning: A commentary. *Developmental Review*, *28*(4), 421–570.

Titmuss, R. M. (1971). *Gift relationship: From human blood to social policy*. New York: Pantheon.

Walton, G. M., and Dweck, C. S. (2009). Solving social problems like a psychologist. *Perspectives on Psychological Science*, *4*(1), 101–102.

CHAPTER 18

Turning Mindless Eating into Healthy Eating

BRIAN WANSINK

Each day, environmental factors such as the visibility, size, and accessibility of food contribute to an ever-growing obesity problem in developed countries. Understanding these drivers of consumption volume has immediate implications for nutrition education and consumer welfare. Yet simply knowing the relationship between environmental factors and consumption will not eliminate its biasing effects on consumers. People are often surprised at how much they consume, and that revelation indicates they may be influenced at a basic or perceptual level of which they are not aware (cf. Langer, 1990; Ross and Nisbett, 1991).

This relates to one of the ironies of consumption and intake research. Although some of the ways environmental factors influence use are intuitively understood by consumers, they often believe that these factors influence other people but not themselves (Pronin and Schmidt, this volume; Pronin, Berger, and Molouki, 2007). Even robust studies involving anchoring, package size, and estimation find that in the debriefings that follow experiments, many consumers believe the results yet vigorously claim those factors had no impact on them (Pronin and Kugler, 2007; Wansink, 2006a).

While the general concepts may not be surprising to people, they are surprised to see those factors have such an impact on them. Furthermore, some researchers and clinicians may be able to predict some of the findings reported here but not be able to offer an explanation of why they occurred. That is, they might be able to predict some of the outcomes without explaining the process.

There are three objectives to this chapter:

1. Illustrate the environmental cues that influence eating behavior and show that they appear to be explained by two basic processes
2. Show that "education" and "awareness" are unlikely to be effective in helping individuals reverse those processes
3. Describe key principles that academics, industry, and government can use when partnering to make perceptible changes in the lives of individuals

Although some academics and policy makers initiate and complete projects believing they will change the world, this may not always be in the way they dreamed. Multiple attempts to do so have suggested some hopeful lessons that could influence both our communication tactics and our leveraging strategies.

Understanding Consumption Quantity and Volume

Consumption is typically studied within a single-period feeding, such as during lunch, during snacks, or during a thirty-minute lab experiment. It is important, however, to realize that total consumption consists of both consumption quantity and consumption frequency. Consuming one chocolate every hour of an eight-hour work day will add to the daily intake as much as comsuming eight chocolates in five minutes. Total consumption intake within a given time period (for instance a 24-hour period) consists of how many occasions a food is eaten (incidence) and how much is eaten during each occasion.

This distinction is important because different factors drive these two variables to differing degrees. The frequency (incidence) that a food is eaten is influenced by the salience of the food and by the effort to obtain and consume it. The volume of food that is consumed in a sitting is influenced by a wide range of other factors and is partly—either knowingly or unknowingly—mediated thorough consumption norms.

Although the bulk of this review will refer primarily to the issue of quantity, or consumption volume, it is important to distinguish between factors that influence how frequently one eats a particular food (the number

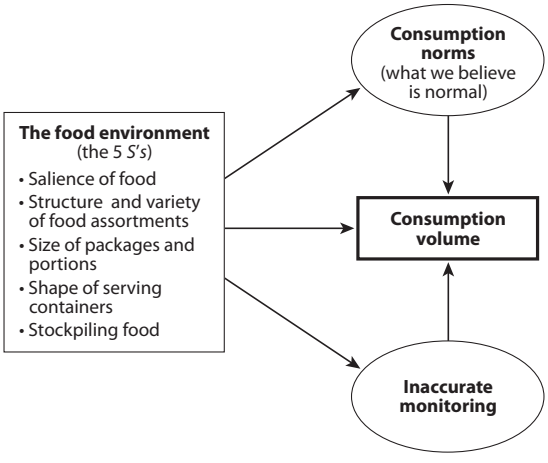
of consumption occasions) versus those that influence how much is eaten on each of these occasions.

The Power of Consumption Norms

People can be very impressionable when it comes to how much they will eat (Herman and Polivy, 1984). Someone can often "make room for more" (Berry, Beatty, and Klesges, 1985; Lowe, 1993) and be influenced by consumption norms around them (fig. 18.1).

For many individuals, determining how much to eat or drink is a mundane and relatively low-involvement behavior that is a nuisance to continually monitor, so they instead rely on consumption norms to help them determine how much they should consume (Wansink and Cheney, 2005). Many seemingly isolated influences of consumption—such as package size, variety, plate size, or the presence of others—may suggest how much is typical, appropriate, or reasonable to eat or drink (Wansink and van Ittersum, 2007). As with normative benchmarks in other situations, these influences may often be relatively automatic and occur outside of conscious awareness (Schwarz, 1996, 1998). Even when consumption norms do influence us, there is anecdotal evidence that people are generally either unaware of their influence or unwilling to acknowledge it (Vartanian, Herman, and Wansink, 2008).

Past evidence of the presence or the absence of this awareness has sometimes been suggested in the context of lab experiments (Nisbett and Wilson, 1977). The problem with trying to generalize from such artificial contexts is that people are generally aware that some manipulation has occurred, and they may be reluctant to acknowledge any influence (Ross and Nisbett, 1991), primarily because of reactance. This phenomenon can best be observed in the context of controlled field studies conducted in natural environments (Meiselman, 1992).



18.1. Antecedents of consumption volume.

Monitoring Consumption Volume

It is well supported that people who accurately monitor how much they eat consume less than those who do not. The trouble is that we are not generally accurate in monitoring how much we eat.

In psychology (Ross and Nisbett, 1991) as well as in food research, there is a robust finding across many studies that people are consistently not aware of the impact that most environmental factors have on them. Even when confronted with large biases that influence consumption by 50% or more, many participants in these studies wrongly maintained that they alone were unaffected. When shown numerical signs in grocery stores that increased purchases by 100%, most claimed it did not influence them even after they loaded eight cans of soup into their shopping cart. The same is true with consumption. That is, when shown that the average person eats 50% more when given a larger container, most people will claim that they would be an exception. When shown that larger plate sizes increased consumption by 25%, they would claim they were an exception.

These perceptions have important implications for people who are trying to be accurate in monitoring and controlling their intake of food. These results underscore that people must take a food's visibility and proximity into account when they try to estimate their prior consumption of it. In general, a food that is more proximate to consume—say cookies on the counter versus those in the cupboard—may be overconsumed relative to what one might think (or recall).

Not surprisingly, a major determinant of how much one eats is often whether he or she deliberately monitors or even pays attention to how much they eat (Polivy and Herman, 2002; Polivy et al., 1986). In lieu of monitoring how much one is eating, people, particularly Americans, will use external cues or rules of thumb (such as eating until a bowl is empty) to gauge the amount of food consumed (Wansink, Payne, and Chandon, 2007).

Unfortunately, using such cues and rules of thumb can yield inaccurate estimates. In one study, unknowing diners were served tomato soup in bowls that were refilled through concealed tubing that ran through the table and into the bottom of the bowls. People eating from these "bottomless" bowls consumed 73% more soup than those eating from normal bowls, but

they estimated that they ate only 4.8 calories more (Wansink, Painter, and North, 2005).

Food-Related Factors that Stimulate Consumption

Outside of hunger itself, perhaps the biggest single driver of consumption incidence is the accessibility of food. Although broadly classified as "accessibility," this can be separated into (1) the salience of the food and (2) the effort necessary to either obtain it or consume it. Whereas the former drives incidence, the later drives both incidence and quantity.

Salience of Food

Studies of what initiates eating episodes among overeaters have reported that one of the main factors contributing to a consumption episode was simply seeing or smelling the food. Consistent with this, a study of eating bouts indicated that 33% of the respondents reported that their most recent eating bout was primarily stimulated by the salience of the food ("It was around or sitting out").

In one study, 30 chocolate kisses were placed on the desks of secretaries either in clear-lidded containers or in opaque ones. Those in the clear containers were more salient and were consumed more quickly (2.9 more each day) than those in the opaque containers (Painter, Wansink, and Heiggelke, 2002; Wansink, Painter, and Lee, 2006). Simply seeing or smelling a favorite food can increase reported hunger (Bossert-Zaudig et al., 1991; Hill, Magson, and Blundell, 1984; Jansen and van den Hout, 1991) and can stimulate salivation (Klajner et al., 1981). There is now even physiological evidence that the sight of food can enhance actual hunger by increasing the release of dopamine, a neurotransmitter associated with pleasure and reward (Volkow et al., 2002).

External cues such as the visual or aromatic prominence of the food can also make a food salient (Schachter, 1971). Given the frequency with which we hear attributions being made to the salience of external cues in everyday conversations, it appears to be an idea worthy of scientific investigation. These externally cued eating bouts are the type that is stimulated when one walks past cookies on the table or sees half of a cake on a counter. In effect, people claim to eat this food "because it's there."

Salience can also be internally generated (Wansink, 1994). It has been found that simply thinking about a food and writing down the different situations in which it has been consumed has been shown to influence how frequently a person intends to eat the product within the next two weeks.

Structure and Variety of Food Assortments

If consumers are offered an assortment with three different flavors of yogurt, they are likely to consume an average of 23% more yogurt than if they are offered only one flavor (Rolls et al., 1981). This example is typical of many consumption situations in which consumers must decide how much of a product to consume when there are no formal guidelines to help them. It has been robustly shown that increasing the variety of a food can increase the consumption volume of that food. It is widely supported that the variety of a food assortment can increase consumption, but the structure of an assortment can also include how this variety is presented (organized vs. disorganized) and how its portions are distributed (symmetrically/equally vs. asymmetrically/unequally).

When there is additional variety, they may consume more because the variety helps reduce satiety, or because there is more likely to be something there that they prefer, or because a consumption norm is suggested by the different flavors that implicitly suggest the amount that is appropriate to consume (consumption rule or benchmark).

In one set of studies by Kahn and Wansink (2004), greater variety (or greater perceived variety) can make a person believe he or she will enjoy the assortment more. Concurrent with this is the notion that the variety or perceived variety can also suggest an appropriate amount to consume. Increasing the actual variety of a product influences its perceived variety, which influences how much one is expected to enjoy eating the assortment, which finally influences consumption volume. Parallel to this, the size of the assortment can also function as a crude consumption norm of what should be eaten in this particular situation.

Recent studies have shown that it may not only be variety that influences consumption, but it may also be the perceived variety one has that drives consumption. One study gave consumers 300 M&M candies that were either in the form of 7 or 10 different colors. Although the candies tasted identical, people who were offered 10 different colors ate 41% more than those given 7 (Kahn and Wansink, 2004). Further evidence of how perceived variety (versus actual variety) can influence consumption was shown when people were offered either organized or disorganized assortments of six flavors of jelly beans. Those offered the disorganized assortment rated the assortment as having more variety, and they took 92% more than those offered the organized assortment.

One explanation that has been offered for this general effect is that it is a response to overcome sensory-specific satiety. Yet this recent evidence suggests that interpretation may not explain all of the variance. Another set of studies involving first and fourth graders replicated the studies by giving children six colored plastic spiders and asking them to take as many as they wanted to play with during the class break. When the colored spiders were disorganized (mixed together) children took 2.1 more spiders than when they were presented to them organized by color. The same was true with colored beads (Kahn and Wansink, 2004).

Size of Packages, Portions, and Inventory

There is a wealth of evidence that packaging and portion sizes have been increasing over the years (Young, 2000; Young and Nestle, 2002). It is well supported that the size of a package can increase consumption (Wansink, 1996), as can the size of a portion in lab environments (Rolls et al., 2004), restaurants (Rolls, Ello-Martin, and Ledikv, 2005), and movie theaters (Wansink and Park, 2001). With packages, the amount by which consumption is raised has generally been shown to be 18%–25% for meal-related foods (such as spaghetti), and 30%–45% for snack foods (Wansink, 1996). In within-subject lab studies, portion size has been shown to increase consumption by 18%–43%, and it increased lasagna consumption by 23% in restaurants and popcorn consumption by 48% in movie theaters.

What is not known, however, is why this occurs. While a "clean your plate" explanation might explain part of this increase (Wansink and van Ittersum, 2007), the same overconsumption is also found when people pour from larger containers of less edible products such as shampoo, cooking oil, detergent, and dog food. Indeed, the package-size studies all used portions (of M&Ms, chips, spaghetti, and so forth) that were large enough such that it was impossible to eat them all in one sitting. Even in these situations, people ate more. Another suggestion, that of the larger packages having lower unit costs, explains 40% of the variation in some situations (Wansink, 1996) but certainly not in situations (such as lab studies) when all the food is given freely.

A general account that is more likely to be operating in all of these situations is one that is more perceptually-based. Recalling the consumption norm notion noted earlier and in figure 18.1, it could simply be that large packages and large portion sizes suggest what an appropriate consumption amount might be. Even if people do not clean their plates or finish the packages, the size of what was presented to them may have given them liberty to perhaps consume past the point that they might have with a smaller, but still unconstrained amount.

What is notable is that package and portion sizes even increase the consumption of unfavorable foods. Moviegoers in Philadelphia were given large or medium-size containers of stale, 14-day-old popcorn. Despite the poor taste of the popcorn, people ate 30% more from the larger size container (Wansink and Kim, 2005).

Stockpiled Food

Bulk buying and warehouse club stores have allegedly contributed to a food-rich environment because they encourage bulk buying. Two of the few studies on this topic compared food-storage habits in homes of obese and nonobese families. Unfortunately, not only did they fail to establish causality, they also failed to find consistent results: the first showed that stockpiled food was more visible in the homes of obese families, but the second showed the opposite (Terry and Beck, 1985).

In one of the few published papers to directly manipulate and observe the impact of stockpiling on consumption, Chandon and Wansink (2002) stockpiled people's homes with either large (12 units) or moderate (4 units) quantities of eight different foods and monitored their consumption for 12 days. They found that when a product was stockpiled, it was used at over twice the rate as the nonstockpiled amount for about the first 8 days after being obtained. After the eighth day, its consumption was similar to that of a less stockpiled product, even though plenty remained in stock.

How does stockpiling a product influence your consumption rate and consumption volume of it? Visibility is one means by which the salience of a food is stimulated (Wansink and Deshpande, 1994), and there is evidence that stockpiling contributes to the salience and visibility of a food, thus increasing how frequently it is eaten and how much is eaten on each occasion. Yet the higher awareness caused by product stockpiling or by point-of-consumption salience is more likely to trigger consideration when the product is convenient to consume than when it requires preparation (Chandon and Wansink, 2002).

Shape of Serving Containers

It has been estimated that approximately 74% of the calories we consume are consumed using intermediate devices such as bowls, plates, glasses, or utensils (Wansink, 1994). How do these items influence our

consumption of foods? We know that the quantity of food presented to a person can increase their consumption volume. What happens, however, if we hold the quantity of food constant and present it to consumers in different size bowls, plates, or serving platters? Earlier work with packaging has shown that even when we give people larger packages that are half full, they eat more than if they are given an equal amount of the product in a smaller package (Wansink, 1996).

Serving oneself a food or a beverage requires cognitive effort to decide what to serve, how to serve, and how much to serve. While part of this process is cognitively driven, part is perceptually driven (Bruner, 1957; Coren and Hoenig, 1972; Gregory, 1972; Neisser, 1967). For instance, if a consumer decides to eat half a bowl of cereal, the size of the bowl acts as a contextual stimulus that may influence how much he or she serves and subsequently consumes. Contextual stimuli, such as the size of a plate or a bowl, may provide signals that consumers use in decision making, even if they are of little diagnostic or rational value (Rao and Monroe, 1988). Such stimuli offer cognitive shortcuts that allow evaluations, decisions, and behaviors to be made with minimal cognitive effort.

Consider glasses. When people look at objects, there is a tendency to focus on height more than width. This tendency to focus on the vertical was noted by Piaget, and it is one reason why people comment on the height of the St. Louis Arch and not on its width (which is the identical size). Studies with teenagers at weight-loss camps showed that people who were randomly given a short, wide 22-ounce glass poured 88% more juice or soda into the glass than those given tall, narrow 22-ounce glasses. They believed, however, that they had poured half as much as they actually did. Similar results were also found with bartenders. When asked to pour 1.5 ounces of liquor into short, wide (tumbler) glasses, they poured 30% more than when instead pouring into tall, narrow glasses (Wansink and van Ittersum, 2003).

Another illusion is the size-contrast illusion. In the context of food, this illusion would suggest that if we spoon 4 ounces of mashed potatoes on a 12-inch plate, we will be more likely to underestimate its size than if we had instead spooned it on to an 8-inch plate. That is, the size-contrast between the potatoes and the plate is greater when the plate is 12 inches than when it is 8 inches.

A study at an ice cream social confirmed this tendency to put more into large bowls. People were given either 12-ounce or 20-ounce bowls and either a tablespoon to scoop their ice cream or a 3-tablespoon scoop. People put 21% more ice cream in the large versus medium bowls, and they dished out 14% more if they used the large scoop rather than the smaller one. With plates and bowls and spoons there is a basic tendency to use their size as an indication of how much should be used. The larger the transfer device, the more we use.

Summary

In the past thirty years, reasonable advances in research and policy have focused on the "outcome" of food intake in different types of environments. The field of food consumption and intake is at a point, however, where the next step needs to be in the direction of understanding the whys behind food intake. The focus needs to explain why we do what we do and not just show it. Doing this would make for more precise research and more effective policies—policies that will have fewer unintended consequences—and will entail more of a focus on developing and testing process-models and theories of consumption. Such models and theories will allow more productive integration across studies and attempts to identify the more fundamental low-involvement drivers of consumption.

Two general mediators that appear to be promising places to start are consumption norms and consumption monitoring. As noted in figure 18.1, both of these are likely to be factors that at least partially mediate the impact of seemingly disparate concepts (such as package size, variety, and social influences) on consumption.

Keeping a focus on the process behind consumption—the whys behind it—will help us move this interdisciplinary field in ways that can raise the profile and the impact of our research on academia, on policy makers, and ultimately on consumers.

Consumption is a context in which understanding fundamental behavior has immediate implications for consumer welfare. Yet simply knowing the relationship between environmental factors and consumption will not eliminate its biasing effects on consumers. People are often surprised at how much they consume, and this indicates they may be influenced at a basic or perceptual level of which they are not aware. The most immediate implication of this research lies in directly altering environmental factors so that they do not have unintended effects. For dieters, diabetics, or those limiting their food intake, the environment can be altered to limit their consumption. Table 18.1 outlines ideas that can serve as further steps in these directions.

Table 18.1 Altering one's personal environment to help reduce food intake

The five S's of the food environment	How one's personal environment can be altered to help reduce consumption
Salience of food: salient food promotes salient hunger	Eliminate the cookie jar or replace it with a fruit bowl. Wrap tempting foods in foil to make them less visible and more forgettable. Place healthier, low-density foods in the front of the refrigerator and the less healthy foods in the back.
Structure and variety of food assortments: structure and perceived variety drives consumption	Avoid multiple bowls of the same food (such as at parties or receptions) because they increase perceptions of variety and stimulate consumption. At buffets and receptions avoid having more than two different foods on the plate at the same time. To discourage others from overconsuming at a high variety environment (such as at a reception or dinner party), arrange foods into organized patterns. Conversely, arrange foods in less-organized patterns to help stimulate consumption in the cafeterias of retirement homes and hospitals.
Size of food packages and portions: the size of packages and portions suggest consumption norms	Repackage foods into smaller containers to suggest smaller consumption norms. Plate smaller dinner portions in advance. Never eat from a package. Always transfer a food to a plate or bowl in order to make portion estimation easier.
Stockpiling of food: stockpiled food is quickly consumed	Out of sight is out of mind. Reduce the visibility of stockpiled foods by moving them to the basement or to a cupboard immediately after they are purchased. Reduce the convenience of stockpiled foods by boxing them up or freezing them. Stockpile healthy, low energy-density foods to stimulate their consumption and to leave less room for their high-density counterparts.
Serving containers: serving containers that are wide or large create consumption illusions	Replace short wide glasses with tall narrow ones. Reduce serving sizes and consumption by using smaller bowls and plates. Use smaller spoons rather than larger ones when serving oneself or when eating from a bowl.

Are Awareness and Education the Solution?

When we review the environmental factors that lead to overeating, there are two consistent explanations for most accounts of overeating: (1) higher consumption norms, and (2) lower consumption monitoring. Yet knowing that these are the two underlying culprits to overeating does not stop us from overeating. First, most people do not believe that these are significant contributors. Second, most people may be unwilling to acknowledge that the environment has any impact on us at all. Although they make over two hundred more decisions each day than they think they make (Wansink and Sobal, 2007), many of these are "automatic" food choices where they unconsciously eat without considering what or how much food they select and consume. This observation is consistent with other psychological work that shows that people tend to have flawed self-assessments, leading to an unmerited overconfidence (Dunning, 2005). With food-intake decisions, their overconfidence may lead to overconsumption and weight gain.

The Problem of Awareness

Consumption occurs within a context where understanding fundamental behavior has immediate implications for consumer welfare. Yet simply knowing

the relationship between environmental factors and consumption will not eliminate its biasing effects on consumers. People are often surprised at how much they consume, and this finding indicates they may be influenced at a basic level of which they are not aware or do not monitor. Our immediate environment can work for us or against us. On one hand, it can unknowingly entice and contribute to our overconsumption of food. On the other hand, it can be less conducive to overeating, leading us to lose weight in a way that does not necessitate the discipline of dieting or the governance of another person.

Given that people so dramatically underestimate the number of food-related decisions they make in a day, it is not unfair to say we often engage in mindless eating. Each of these small decisions is a point at which a person can be unknowingly influenced by environmental cues. In the interest of better controlling food intake, people need to be more aware of the number of decisions that influence what they eat, as well as when they start eating and when they stop.

One study (Wansink and Sobal, 2007) suggested that we make over 200 more food-related decisions each day than most of us realize. Each of these decisions that we are not consciously aware of provides an opportunity for being unknowingly influenced by environmental cues. Here I will show that people (1) are not aware of overconsuming, or (2) not aware of being impacted by these cues *after* the cues and their general impact are made salient.

A Meta-Analysis of Awareness

Consider four controlled field studies that investigated how environmental factors such as package size, serving bowl size, and plate size influenced how much people consumed in natural environments when the people were randomly assigned to an exaggerated treatment condition. Participants in these studies spanned a wide range of ages and backgrounds (including graduate students, moviegoers, and Parent Teacher Association members), and in each study they were systematically assigned to different conditions and their consumption behavior was assessed. Across all of these studies, the same two questions were asked of those in the exaggerated (e.g., big bowl) treatment conditions:

1. "How much did you eat compared to what is typical for you?"
2. "In this study, you were in a group that was given [a larger bowl]. Those people in your group ate an average of 20%–50% more than those who were instead given [a smaller bowl]. Why do you think you might have eaten more?"

The qualitative data collected during the postexperiment debriefings were coded using content-analysis procedures (Neuendorf, 2002; Webber, 1989). The answers to the first question about amount eaten were coded as either "less than," "about the same," or "more than." The second question about explanations for overeating was coded as (1) they denied eating more, (2) they attributed it to hunger, (3) they attributed it to the intervention, or (4) another explanation (being in an exciting situation, etc.). Individual calculations of coding reliability between the two coders were alpha = .94 (for the "how much" question) and alpha = .74 (for the "why" question). Much of the variability for the why question was due to the answers that were subsequently coded upon agreement as "miscellaneous."

In total, 379 people were involved in these field studies, with 51% (192) being in the exaggerated environmental-cue condition. Brief descriptions and results for each study are shown in table 18.2. Within these treatment groups, the average increase in consumption over the control was 31%. However, an average of 73% of the participants believed they ate as much as they normally ate. Of those remaining, an average of twice as many believed they had eaten less than those who thought they might have eaten more (19% vs. 8%). For those 8% to have eaten enough to fully account for this 31% increase, each would have had to eat an average of 387% more than the average member in the control group.

When told of their treatment groups' bias and when asked why they might have eaten more, 52% claimed they did not eat more, and 31% said that if they did eat more, it was because they were hungry. Only 2% of the participants believed they had eaten more because of the environmental cue that had been specifically named. Fifteen percent claimed they ate more for miscellaneous reasons, such as because it was a special occasion (the Super Bowl) or because it was "free."

Of those who did believe it possible that they ate more, only 2% acknowledged it was because of the environmental cue. The hesitancy to acknowledge being influenced by an external cue is common and has even been found when people are presented with tangible evidence of their bias. For instance, when pouring a standard drink of alcohol, the horizontal-vertical illusion has lead professional bartenders with over five years of experience to pour an average of 29% more alcohol in short, wide glasses (tumblers) than tall, narrow glasses (highball glasses) which held the same volume (Wansink and van Ittersum, 2005). When confronted with their bias and when shown that they poured at average of 1.9-ounces compared to the 1.5-ounces that was prescribed, the general

Table 18.2 Field-study participants deny the influence interventions have on their intake behavior

Sample and context of study	Intervention and findings	"How much did you eat compared to what is typical for you?"				"In this study, you were in a group that was given [a larger container]. Those people in your group ate an average of 20–50% more than the others. Why do you think you might have eaten more?"[1]				
		Less (%)	About the same (%)	More (%)	Chi-square	"I didn't eat more" (%)	"I was hungry" (%)	"The (intervention) influenced me" (%)	Other (%)	Chi-square[2]
40 MBA students at a Super Bowl party in a bar in Champaign, IL (Wansink and Cheney, 2005)	Those serving themselves Chex Mix from 4-liter bowls ($n = 19$) served 53% more than those serving from 2-liter bowls	23	57	20	10.55 ($p < .01$)	63	31	3	3	22.78 ($p < .001$)
98 adults preparing a spaghetti dinner for two in Hanover, NH (Wansink, 1996)	Those given half-full 32-oz boxes of spaghetti ($n = 51$) prepared 29% more than those given full 16-oz boxes.[3]	18	73	9	70.36 ($p < .001$)	71	27	4	8	67.76 ($p < .001$)
161 afternoon moviegoers in a Chicago suburb (Wansink and Park, 2001)	Those given 240-g buckets ($n = 82$) ate 53% more than those given 120-g buckets	19	75	6	128.77 ($p < .001$)	15	77	5	3	152.00 ($p < .001$)
158 evening moviegoers in Feasterville, PA (Wansink and Kim, 2005)	Even when given stale, 14-day-old popcorn, those given 240-g popcorn buckets ($n = 40$) ate 34% more than those given 120-g buckets of the same popcorn	14	78	8	141.65 ($p < .001$)	12	79	2	7	179.42 ($p < .001$)
Average across all studies (Weighted by the number of subjects per study)		19	73	8	331.26 ($p < .001$)	52	31	2	15	203.97 ($p < .001$)

Note: Answers are from those in the treatment group who received the intervention that resulted in greater consumption.

[1] The specific intervention in the study was noted at this point. Here, the example of larger bowls was used.

[2] The chi-square test was conservatively conducted excluding the "Other" response from the analysis. Including this resulted in all *p's* < .00.

[3] In this study, people poured spaghetti but did not actually consume it. Questions were modified to reflect pouring instead of eating.

reaction was one of disbelief and denial, despite the tangible evidence (Wansink and van Ittersum, 2003).

Why Education and Awareness Do Not Seem to Work

Lab studies have often found that people either do not believe they were influenced by external cues or do not want to admit that this was the case (Nisbett and Wilson, 1977). While such studies have not been systematically evaluated, their anecdotal evidence has often been discounted because of their demand effects (Vartanian, Herman, and Wansink, 2008). Using field studies, I have shown here that people claim to be unaware of these factors, ultimately increasing their consumption. Even when confronted with empirical data, most participants in environmental manipulations continue to disavow the findings or to look for alternative explanations. Although these results do not fully disentangle unawareness from denial, the consistency of the findings across studies points to a strong systematic influence that goes beyond what people either know or will confess.

Regardless of the reasons that researchers do not always see the change they hoped to influence, what remains is that they may have much less of a direct impact on people than they might want to believe. Even when they do happen to be successful in generating awareness of their work—be it through publications, presentations, or press—it is not clear that it is presented in a way that indicates to people how they should change (see figure 18.2). As a result, people can hear something "interesting" but be uncompelled to make a decision to change in their own life.

Even when people do wish to change—which can be the case when a person takes away diet tips from a new research study—something prevents them from doing so. In some cases it may be the inertia or structural barriers in their life. In other cases, it simply may be that we do not provide them the support structure—the choice architecture (Thaler, Sunstein, and Balz, this volume)—that is needed to make the difference between a nice idea and a nice change.

Turning Mindless Eating into Healthy Eating

There are two common levels of analysis within the large ecological context of the food environment: a macro level and a micro level. At the macro level, the focus is on government regulation, food-industry incentives, school lunch programs, and advertising campaigns (Brownell and Horgen, 2003). At the micro level, the focus is on making a choice—such as that between fresh fruit or a sweet snack.

Within this broad ecological context, there is an intermediate level that is often overlooked because it lies between the arenas of policy and personal choice. This intermediate level is the environment in which we live and work. It is a level that can influence food intake without involving the taste, texture, or quality of the food itself. That is, regardless of whether one is eating an apple or an apple pie, environmental factors can often unknowingly drive intake. To avoid having to continually make caveats about different food categories, it is useful to differentiate those drivers that are independent of the food from those that are more dependent.

High at the 30,000 foot-level, critics blame low-priced, easily available food for making us fat. Some blame government subsidies to agriculture, supersizing food companies, and even the schools. If all of these were gone, our environment would clearly be less obesigenic. Would we all revert back to having the sleek, trim figures of people we see in 1950s black and white photos? That is less clear. Changing capitalism and changing the world is a slow process. In this case, it is not clear how much of the world wants to change, or even how trim it would actually make us (see fig. 18.3).

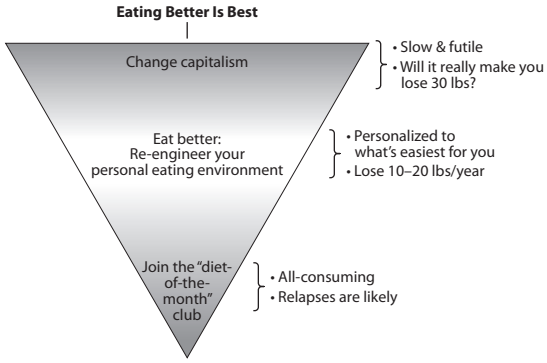
Neither this micro nor macro approach holds bright promise for the person who wants to get his or her family or him or herself back on the right track. One extreme is slow, futile, and unlikely to work; the other is all consuming and prone to relapse.

The key to change may lie more in the middle. People can reengineer their personal food environment to help them and their families eat better. They can move from mindless overeating to mindless *better* eating.

The Laboratory of Life

Academics and policy makers often observe inconsistencies in the choices made by people. Some researchers specialize in uncovering the biases that lead to poor choices. But knowing what people should do does not make the researchers and policy makers

18.2. Making people "aware" is not enough. We need to help them act and change.



18.3. Where is the greatest potential for change?

experts on helping those people change—telling them is not always helping them. In some ways, the researchers and policy makers are as unrealistically naive about their power and influence as the participants who believe they are not influenced by the size of a bowl or the "health halo" from a menu.

Much academic research collects data in highly controlled behavioral laboratories, where reasoned thought is king and decisions are made on a computer keyboard. Most of the professional hours of policy makers is spent in chaotic, tension-filled rooms with special interests and competing positions, where the realities of human behavior are replaced with positions or assumptions about what is best and what people ought to do.

It is easy for them to believe they are on the right track if they look at the wrong cues. Academics can believe their ideas are having an influence if they are accepted by good journals and cited by colleagues. Policy makers can believe their ideas are having an influence if they receive funding or they result in laws that are passed.

What is missing from many of their lives as academics or as policy makers is the experience they can gain from the laboratory of life. They only know whether their ideas have been accepted by other academics and policy makers, not whether they have been accepted by the people who were originally intended to be helped.

The National Mindless Eating Challenge

Because people are generally unwilling to believe that their eating habits are impacted by the environmental cues around them, they respond in two ways. Either they are resistant to general advice, or they do not know where to start to make the stylized changes that would be most useful for someone in their particular situation.

To help provide a solution to this dilemma, the National Mindless Eating Challenge was formed. From a research perspective, there were three initial objectives of the challenge:

1. Track the field success of lab-tested tips
2. Develop profiles of dieter typologies to provide stylized feedback
3. Determine what form of feedback best encourages compliance and success

After making preliminary discoveries, we should be able to use these insights to help close the gaps in figure 18.2. Doing so would tighten the link between communicating insights and actually changing behavior.

METHOD

More than 10,000 individuals who were interested in changing their eating behavior registered on a website (www.MindlessEating.org). A random sample of 2,500 of these individuals were offered the opportunity to be involved in the first stages of calibration for the National Mindless Eating Challenge. While 17% of these had read the book, *Mindless Eating: Why We Eat More Than We Think* (Wansink, 2006), the majority had not.

The book had identified five common eating goals (1–5 below). Based on interactions with people after the book was published, four additional goals were added (6–9). The initial segmentation of individuals was based simply on which goal they wanted to focus on in the upcoming month. The nine eating goals were

1. Reduce meal stuffing
2. Reduce snack grazing
3. Reduce restaurant indulging
4. Reduce party binging
5. Reduce desktop/dashboard dining
6. Eat more
7. Eat better
8. Help family eat less or healthier
9. Maintain weight loss

Participants were then asked to complete a lengthy online questionnaire that included self-reported measures of wellness, weight, productivity, happiness, and recent medical history. Following this, they were randomly given three food-behavior suggestions from a pool of tips that had been empirically supported by academic research. For each of the nine eating goals, 8 to 21 lab-tested tips were identified as being potentially relevant.

The basic goal of the program was to provide small, easy-to-implement changes that could be made painlessly but that would yield sizable results over time.

Based on the estimated calorie savings of each change, the average person who was successful in implementing half of their changes would lose the mathematical equivalent of 15-20 pounds a year if everything else remained relatively equal in their lives (including their metabolism).

After each individual in the challenge disclosed his or her primary eating goal for the month, the individuals were randomly given three of the suggestions from the relevant pool. They were then asked to write down (1) their biggest barrier to implementing that tip, (2) one strategy they could use to overcome that barrier, and (3) an estimate of how many days in the next thirty they would be able to successfully accomplish that behavior.

They were then provided a Monthly Margin Daily Checklist to track their progress (see figure 18.4 for example). They were instructed to write their three changes for that month onto the three rows of the checklist and to check off each day they successfully made the change.

Each Friday for the next four weeks they received a reminder through email, and they were also given a range of the amount of weight they would lose over a year based on the number of checks they had made. At the end of the month, they were invited back to the website where the process repeated itself.

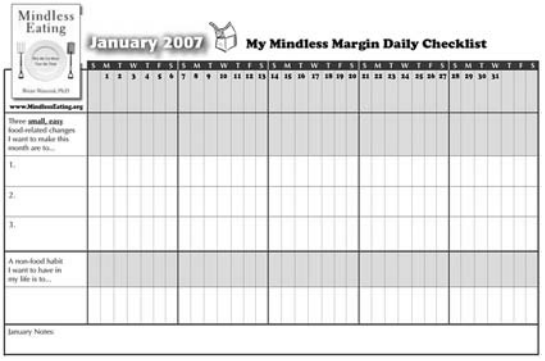
The process was repeated for three months, after which changes in wellness, weight, and compliance were assessed for each of the tips in each of the goal groups. A composite measure of wellness was developed based on their answers to happiness and wellness questions. Compliance and weight change were measured separately (Wansink, Just, and Payne, 2009).

RESULTS

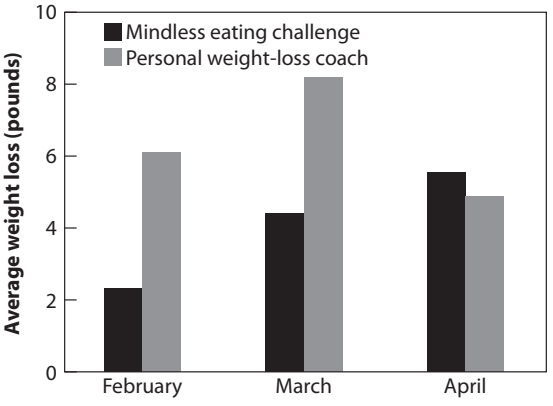
Among the 2,374 who completed a month or more of the program, the average weight loss was 2.3 pounds (with a range of +0.8 to –12.6 pounds). Preliminary analysis of those who stayed in the program for all three months (1,743) showed they achieved an average weight loss of 5.2 pounds. As a point of comparison, the same month the challenge was started, another program was initiated—a traditional face-to-face program with 73 individuals who were given two months of free dietary advice and free access to the health club (Personal Weight Loss Coach). Their reported weight loss over the first two months was 8.2 pounds but had dropped to 4.9 pounds at the end of the third month.

The reported weight of those involved in the Mindless Eating Challenge gradually continued to decrease over each of the three months (–2.3, –4.4, –5.5 pounds). In contrast, the weight loss of those in the Personal Weight Loss Coach program was initially more drastic (see figure 18.5) but the rate declined after the two-month program was over (–6.1, –8.2, –4.9 lbs.). In spite all of the advantages of the "high touch" Personal Weight Loss Coach, the trend of progress after the second month became concerning since the weight changes were not as apparent. In contrast, the slow-and-steady mindless changes suggested by the Mindless Eating Challenge continued to show increased effectiveness with the passage of each month.

It must be understood that these were not randomly selected groups of people. Those people who came to a website or those who signed up for a personal trainer may not be comparable along certain dimensions. For instance, some of these might



18.4. Illustration of an accountability chart used with the National Mindless Eating Challenge.



18.5. The National Mindless Eating Challenge led to more steady weight loss than a two-month personal weight-loss coach.

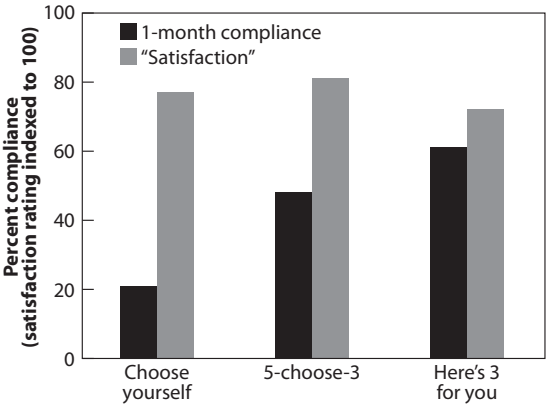
have been restrained eaters (Ward and Mann, 2000), and others might have struggled with ego depletion (Baumeister, 2002). Additionally, this study examined only those who stayed with the programs. A larger group of people dropped out of the web-based program than out of the personal trainer program. That is why it is important to condition these results by saying among those completing three months of the program, steadier weight loss was found among those who had been involved in the web-based program than in the personal coach program.

While segmenting individuals based on their eating goals was more effective than giving general advice, further segmentation strategies were explored using the various measures that had been obtained. A promising segmentation strategy that had proven useful in post hoc analyses involved two dimensions of control: (1) the general degree of self-control one exhibited or claimed to have, and (2) the degree of control one had over their immediate environment. Whereas the first dimension is self-explanatory, the second is intended to distinguish between an individual who can easily make changes to his or her environment (perhaps someone who lives alone or otherwise controls food purchasing and preparation decisions within a household) from someone who instead is a member in a large household. Making dichotomous splits between these two variables yielded four behavioral segments (low/low, low/high, high/low, high/high). When multiplied by the nine different goals, this resulted in 36 reasonably homogeneous segments. A further analysis of the success of various lab-tested tips within these 36 segments yielded a rank ordering of what appeared to be most effective for these different profiles of individuals.

In a series of follow-up studies, we experimented with how to most effectively present these suggestions to individuals in a way that would encourage the greatest compliance and satisfaction. One study worth noting involved 3,000 people who where presented these tips in one of three ways:

1. The Choose Yourself group: Here are 173 lab-tested tips. Choose 3 to use this upcoming month.
2. The Five-Choose-Three group: Here are 5 tips most statistically related to success for someone like you. Choose 3 to use this upcoming month.
3. The Here's Three group: Here are 3 weight-loss tips most statistically related to success for someone like you.

All participants were told they could also make up their own tip and that they did not need to be constrained to what we suggested. One example of this that led a couple to lose 33 and 44 pounds was their self-generated tip to not sit down for lunch or dinner unless both a fruit and a vegetable were on the table. Across all three groups no more than 4% generated their own tips.



18.6. Libertarian paternalism and dietary choice: less choice appears more effective.

For one month we tracked compliance (the percentage of tips implemented each day) and satisfaction. We found that giving a person too much latitude in their choices reduced their compliance. Giving people less choice did not lead to low satisfaction. As can be seen in figure 18.6, satisfaction was still rated in excess of 70%. The libertarian paternalistic idea of reducing the number of defaults (3 suggested choices instead of 173) but still allowing free choice ("make your own tip") led to more compliance and higher satisfaction. It is still unclear what the long-term progress would be, and further study is needed across time and different groups of people.

Four Thoughts about Changing Eating Habits

This laboratory-of-life experience—trying to influence behavior in the real world—brings lessons of both discouragement and encouragement. Its results are discouraging because they show that even the most practical insights are difficult for people to implement because they do not recognize their relevance, they lack motivation to make them work, or they lack the step-by-step encouragement and direction.

The encouragement that comes from the laboratory of life is that there is an easy way to be productive in both ivory towers and public policy bunkers while also taking a break to breathe fresh air that inspires fresh thinking. In working with the National Mindless Eating Challenge, here are four lessons I have learned that are now my working hypotheses on how to help people translate our insights into action.

PROVIDE EVIDENCE THAT THE CHANGE WILL WORK

The world is filled with advice. Our forte as academics is our ability to prove or disprove the effectiveness of our ideas and programs. We can prove that, on average, people eat 27% more when given a 12.5-inch plate than a 10.5-plate. If a dietician were to command a person to use smaller plates, it might engender reactance. If we say it with proof, we can engage reason.

All of the changes suggested in the National Mindless Eating Challenge have empirical proof that they influenced single-session intake by at least 12%. While we do not always know how this would translate into a person eating at a truckstop in Bristow, Oklahoma, or eating at Thanksgiving in Correctionville, Iowa, we have good reason to believe they would still be more likely to work than not.

GIVE A STYLIZED SET OF CHANGES

As the grand inquisitor scene in *The Brothers Karamozov* suggested, too much freedom of choice can be paralyzing to many people. Recall that in the National Mindless Eating Challenge one group of people was told they could choose whatever changes they wanted. Yet compliance to those changes was not high in this group. Instead, compliance was highest—exceeding 70%—when we told them *what* changes were most correlated with success for someone like them and did not allow them to select their own. There are a couple important elements to understand. First, they were told specifically what to do. Second, they were given evidence that this was not generic advice but rather what was uniquely relevant to them.

Where is the freedom of choice in a situation such as this? It is the escape hatch that they are still instructed that they can choose any other tip they wish. Although they might appreciate the freedom of having the option, less than 4% ever took the option, and these were often hold-over tips from a prior month.

GIVE A TOOL FOR DAILY PERSONAL ACCOUNTABILITY

One often-mentioned rule of thumb in behavioral modification says it takes about 28 days—one month—to break an old habit and to replace it with a good one.

At the end of every day, people were asked to check off the changes they accomplished that day (fig. 18.7). This small act of accountability is intended to make people more mindful of what they are doing, more accountable, and provides its own small reward.

As mentally disciplined as most of us like to think we are, nothing beats having to face facts each night and check off a little box. We have very selective memories, but the Power of Three checklist lets us know just why—or why not—we have painlessly lost two pounds on the thirty-first of the month.

March	1	2	3	4	5	6	7	8	9	10	11	...	31	Total
Use the 1/2 plate rule - 1/2 veggies or salad	X	X	X		X	X	X	X	X		X	...	X	27
Slow down - start last; finish last		X			X					X	X	...	X	13
Only serve vegetables "family style"	X	X	X		X	X	X		X	X	X	...	X	24

18.7. The Power of Three checklist.

GIVE REGULAR ENCOURAGEMENT AND FEEDBACK

Habits are reinforced by days of scripted behaviors. Suggested behavioral changes—regardless of how compelling, are likely not to work when they encounter the tyranny of the moment.

Providing some sort of community of encouragement can help move behavior changes from experiments to habits. With the National Mindless Eating Challenge, there are three major ways we try to provide encouragement and a sense of a supportive virtual community.

First, we provide weekly reminders and encouragement, telling people how much weight they would lose over a year based on how well they have done that week on making their behavior changes. Second, we reengage with them at the end of every month to hear how well things have gone for them and to provide fresh suggestions. Third, we occasionally share ideas or solicit their feedback on various topics we think would be of interest to them.

Partners in Change: From Individuals to Industry to Government

As academics we have the flexibility and unfettered creativity to develop ideas, lab-test these ideas, and even concept test them in the laboratory of life. To scale up their influence, it can help to have a partner. In the case of health-care, both companies and governments have much to gain.

From Proof of Concept to Proof of Profitability: A Retailer Case Study

Taking ideas from tested to *practicable* is important. What could accelerate their adoption, however, is taking ideas from tested to *profitable*. Given the costs of health care, the corporate incentive to encourage healthier eating may be changing.

One way to move from proof of concept to proof of profitability is in a corporate setting. As part of the

retailer's benefits package, a number of employees from individual stores were involved in the National Mindless Eating Challenge, which was launched on November 26, 2007. All employees of these stores were given the option of signing up for the program at no cost. After signing up and providing self-reported measures of well-being, they were asked questions that enabled them to be segmented into one of 36 profiles based on their eating goal, their self-control, and the extent they could control the food environment in their home.

Following their completion of the survey, they were given table 18.3 which emphasized the basic principles of the program. They were then given a Power of Three checklist (recall fig. 18.7). They were encouraged to write in the three small changes they had been given to try each day for the next month and to place their monthly checklist next to their bathroom mirror or their bedside and to complete it each evening.

On each Friday, participants were sent a reminder to keep up with their checklist and some words of encouragement. They were also given the opportunity to go to a specified web address to report the total number of days they successfully made each of their three behavior changes (0–21). Based on the number of successful changes they had made and based on our estimated range of calorie savings per change (calibrated by lab studies), we were then able to give them automated feedback of a range of how many pounds they would lose over the course of a year.

At the end of each month, the employees were invited to return to the website and report their experience. They could change their eating goal at that time. Based on a new eating goal and based on the success of their prior month, they would then be given three new suggested changes to make the next month. Should they have wished to use one of their own changes, they could also do so.

A company such as these retailers is interested in having happy, healthy, productive employees. If an employee can move to a healthier weight, it could have implications for both health-care costs and productivity.

While the payback to this program has not yet been determined, the key lesson from this brief description of a case study is not in the details of the success of what the program might show in terms of reduced health-care costs and increased productivity. The lesson is that it may not always be an impossible step from an interesting insight to an impactful one. It may also be a step that is worth taking with the type of partners—corporations—that we would not otherwise have considered.

The Government for the People and for Better Nutrition

All governments are filled with examples of tremendously cost-effective programs that have transformed the lives of people involved. Unfortunately, these governments also have examples of programs that have been disappointingly ineffective. In fact, there

Table 18.3 Mindless eating: Developing your own plan to mindlessly eat better

Key points	Advantages	Disadvantages
Your mindless margin. By making 100–200 calorie changes in our daily intake, you won't feel deprived and backslide. **Mindless better eating**. Focus on reengineering small behaviors that will move you from mindless overeating to mindless better eating. Five common places to look (diet danger zones) include meals, snacks, parties, restaurants, and your desk or dashboard. **Mindful Reengineering.** To trim your mindless margin, you can use basic diet tips, but a more personalized approach is to use (1) Food Trade-offs, or (2) Food policies. Both give you a chance to eat some of what you want without making it a belabored decision. **The Power of Three**. Design three easy doable changes that you can mindlessly make without much sacrifice. **Mindless margin checklist.** Use this daily checklist to help you move from mindless overeating to mindless better eating.	Easy and inexpensive No hunger or deprivation Easy to use with family members No foods are off-limits, but portions are reduced Can fit any routine; flexible to what a person thinks will be easiest for him or her Weight stays off Can be used instead of other diets, along with other diets, or after other diets	Weight loss is gradual: a 100–200 calories a day is 10–20 lbs a year. Until small changes become second nature, this works best when a daily habit checklist is used.

are perhaps as many reasons for failure as there are programs. One idea this chapter has continually emphasized, however, is that the presumption that education and awareness will change behavior is both wrong and insufficient.

CENTER FOR NUTRITION POLICY AND PROMOTION

The U. S. Department of Agriculture (USDA) is the standard bearer of health promotion as it relates to nutrition. Responsible for providing the USDA guidance is the Center for Nutrition Policy and Promotion (CNPP). It establishes new dietary guidelines every five years and its website (formerly MyPyramid.gov) is the second most visited government website aside from the IRS. It provides the nutritional guidance for decisions made about the WIC program, the Food Stamp program, and the School Lunch program.

In 2007, I received a presidential appointment to be the Executive Director of CNPP for fifteen months, until the next administration came to office (Squires, 2008). The first change made was to refocus the efforts of the center on where it could make the most difference most immediately. Because the "nutritional gatekeeper" is estimated to influence around 72% of the eating decisions made by his or her family members (Wansink, 2006b), all CNPP efforts were refocused on influencing this person. The overall objective was to aim at 24/7 360-degree nutritional information. We wanted to touch or nudge these gatekeepers wherever they purchased and prepared food and wherever they worked and played (IFIC, 2008):

1. Messaging was retargeted and new media—such as podcasts, YouTube, audio clips, online games, interactive tools, and home activities—were designed and launched.
2. Four new tools—the MyPyramid Menu Planner, MyPyramid for Pregnant and Breastfeeding Mothers, MyPyramid for Preschoolers, and the Cost-of-Feeding a Child Calculator—were launched.
3. Over 100 corporate partners were enlisted through the Partnering with MyPyramid program, which incentivized them to promote the dietary guidelines however and wherever they wanted—on packaging, online, in stores, schools, homes, and so on.

Part of the insights the CNPP used in developing new tools and approaches for outreach involved understanding what ideas had been effectively disseminated at the local level and what ideas had been effectively disseminated by corporations. As a result of using these ideas, web hits to the MyPyramid.gov website increased 44% in 15 months, up to 5.6 million hits per day, making it the most accessed federal (.gov) website (Wansink, 2009).

Influencing the nutrition tools that agencies like the CNPP use involves "life evidence" as much as it involves lab evidence. This need raises the importance of finding the most convincing context in which to develop a proof of concept. Sometimes that path might be through a grant from the National Institutes of Health, whereas in other cases it could be through the success of a retailer's benefit package.

STATE EXTENSION AND FRONTLINE NUTRITION

Many of those who most desperately need nutrition guidance are those who can afford it least. Fortunately, many pockets of need are covered by frontline programs sponsored by state extension programs and other personnel (such as WIC or EPNEF).

Yet while these programs have contact time, this does not mean they are as effective as they could be. The four findings noted above would be ones that could very easily be used to help these educators slightly modify their approach in a way that might have longer-term influences. Here is a review of the insights from our experiences in outreach:

1. Provide evidence the change will work.
2. Provide a stylized set of changes.
3. Provide a tool for daily personal accountability.
4. Give regular encouragement and feedback.

In working with various educators from these programs, I have observed that many are very good at providing regular encouragement and feedback. This quality is facilitated by the fact that some programs, such as WIC and EPNEF, necessitate regular meetings and feedback sessions.

What is not consistently evident is the use of the first three principles. Too often an overworked educator will give a person a list of eating commandments. These are sometimes provided with little convincing evidence the methods will work, with little regard for their personal circumstances, and with little guidance on how to regularly stay the course.

While effective web-based algorithms can stylize feedback and provide continued support, it is wrong to believe that a similar, but "higher-touch," approach would not work for frontline nutrition staff. A basic flowchart offers a reasonably stylized amount of feedback for the most common profiles. Branching questions that separate people into diagnostic profiles could then be used to provide more relevant and proven suggestions. Furthermore, free resources, such as a printed version of the monthly checklist, could be given as tracking tools. Individuals can use such resources to track themselves and to show educators

so they know how to adjust what they teach and the advice they give.

No Programs without Partners

Perhaps the most effective way to induce change is to have somebody else help. As policy makers or academics, what we can do by ourselves or with our own agency is limited. Regardless of how talented and experienced we believe we might be, we are limited in our vision, our resources, our connections, and so on. Policy makers and academics become much less limited, however, when they partner with others whose strengths can match the former's weaknesses.

One of the most effective rules of thumb we could use when thinking of new ideas and new outreach efforts is: no programs without partners. Our partner can be a Fortune 500 food company or a school lunch program in Meridan, Ohio. It can be another agency or another researcher. It can be a benefits officer or a journalist.

Consider two projects that benefited from cooperation with multiple partners: (1) the Smarter Lunchrooms Project and (2) the Small Plate Movement.

The Smarter Lunchrooms Project (www.SmarterLunchrooms.org) is aimed at improving the food choices students make at school. Instead of restricting the types of foods that are available, the goal is to modify school lunchrooms in a way that guides or nudges students toward better choices without them realizing it. The notion of constrained volition (Wansink, Just, and Payne, 2009) involves altering an environment in a way that gives people the illusion of choice by constraining them in an imperceptible manner. Studies have shown that students ate more healthily when food was reorganized and trays were eliminated and when they had to pay cash (versus a debit card) for desserts and sugared drinks (Wansink, Just, and Payne, 2009).

Two partners were critical in making this happen. The USDA sponsored some of the earlier studies through a cooperative agreement. After the studies were conducted, a series of joint research bulletins were produced with USDA research sponsors. Following this, the School Nutrition Association was enlisted to help disseminate the findings through their annual conference, website, and mailings to members. While the project could have been conducted without the cooperation of either group, their partnerships helped sharpen and disseminate the message to more groups with more effectiveness.

The second illustration is the Small Plate Movement. Given the encouraging results reported earlier about using smaller plates, the Small Plate Movement was launched (www.SmallPlateMovement.org). The purpose was to help consumers make a small change (use smaller plates) that could help them eat less and have a possible ripple throughout other parts of their lives. By signing the Small Page Challenge, they agreed to use a plate no larger than 10 inches for at least a month.

There were five key partners. The first was the Cornell University Food and Brand Lab, which developed the website and the campaign. The second was the TOPS (Take Off Pounds Sensibly) weight-loss group, which spread the word through group meetings and the blogs and homepages of individual members. The third was the National Restaurant Association, which promoted the use of smaller plates among members citing reduced food costs and higher perceptions of value. The fourth were plate manufacturers and importers (such as MindlessProducts.com). And the fifth was an 18,000 person town in southern Minnesota.

Albert Lea, Minnesota was selected as a proof of concept of how mindless-eating solutions could be put into place in real life. As the eating component of the Blue Zones Vitality Project, the principles from the book *Mindless Eating* became part of a pledge taken by over 2,000 families. One of the changes people could make was to use a smaller (less than 10-inch) dinner plate for six months. Although many other changes were also being made (including physical activity), smaller plates were one component whose anticipated success was bound to result in a win-win partnership.

If we cannot find another person or organization to partner with us and our program, there could be two problems: (1) our program is not as clear and compelling as it could be, or (2) it could be a bad idea. In either case, trying to enlist a partner will make even the early stages of program development more efficient. While this means sharing the credit, most people would think it better to share the credit for a success than the alternative.

Conclusion

The nineteenth century has been called the century of hygiene. That is, in that century more lives were saved or extended as the result of an improved understanding of hygiene and public health more than by any other single method. The twentieth century was the century of medicine. Vaccines, antibiotics, transfusions, and chemotherapy all helped to contribute to longer, healthier lives. In 1900, the life expectancy of an American was 49 years. In 2000, it was 77 years.

I believe the twenty-first century will be the century of behavior change. Medicine is still making

fundamental discoveries that can extend lives, but changing everyday, long-term behavior is the key to adding years and quality to our lives. This factor will involve reducing risky behavior and making changes in exercise and nutrition. The more we exercise and the better we eat, the longer and more productively we will live. There is not a prescription that can be written for such behavior. Eating better and exercising more are decisions we need to be motivated to make.

When it comes to contributing the most to the life span and quality of life in the next couple of generations, behavioral scientists could be well suited to effectively help us make the move and get both of these done. And why not start with our eating habits?

References

Baron, R. M. and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical consideration. *Journal of Personality and Social Psychology*, *51*, 1173–1182.

Baumeister, R. F. (2002). Yielding to temptation: Self-control failure, impulsive purchasing, and consumer behavior. *Journal of Consumer Research*, *28*, 670–676.

Berry, S. L., Beatty, W. W., and Klesges, R. C. (1985). Sensory and social influences on ice-cream consumption by males and females in a laboratory setting. *Appetite*, *6*, 41–45.

Birch, L. L., and Fisher, J. O. (2000). Mother's child-feeding practices influence daughters' eating and weight. *American Journal of Clinical Nutrition*, *71*, 1054–1061.

Birch, L. L., McPhee, L., Shoba, B. C., Steinberg, L., and Krehbiel, R. (1987). Clean up your plate: Effects of child feeding practices on the conditioning of meal size. *Learning and Motivation*, *18*, 301–317.

Bossert-Zaudig, S., Laessle, R., Meiller, C., Ellgring, H., and Pirke, K. M. (1991). Hunger and appetite during visual perception of food in eating disorders. *European Psychiatry*, *6*, 237–242.

Bradburn, N., and Sudman, S. (1981). *Asking questions.* San Francisco: Jossey Bass.

Brownell, K. D., and Horgen, K. B. (2003). *Food fight: The inside story of the food industry, America's obesity crisis, and what we can do about it.* New York: McGraw-Hill/Contemporary Books.

Bruner, J. S. (1957). On perceptual readiness. *Psychological Review*, *64*(2), 123–152.

Chandon, P., and Wansink, B. (2002). When are stockpiled products consumed faster? A convenience-salience framework of post-purchase consumption incidence and quantity. *Journal of Marketing Research*, *39*, 321–335.

———. (2007). Is obesity caused by calorie underestimation? A psychophysical model of fast-food meal size estimation. *Journal of Marketing Research*, *44*(1), 84–99.

Clendennen, V., Herman, C. P., and Polivy, J. (1994). Social facilitation of eating among friends and strangers. *Appetite*, *23*, 1–13.

Coren, S. and Hoenig, P. (1972). Effect of non-target stimuli upon length of voluntary saccades. *Perceptual and Motor Skills*, *34*, 499–508.

De Castro, J. M. (1994). Family and friends produce greater social facilitation of food-intake than other companions. *Physiology and Behavior*, *56*, 445–455.

Dunning, D. (2005). *Self-insight: Roadblocks and detours on the path to knowing thyself.* New York: Psychology Press.

Evans, G. W., and Lepore, S. J. (1997). Moderating and mediating processing in environment-behavior research. In G. T. Moore and R. W. Marans (Eds.), *Advances in Environment, Behavior and Design* (Vol. 4, pp. 256–286). New York: Plenum.

Fisher, J. O., Rolls, B. J., and Birch, L. L. (2003). Children's bite size and intake of an entree are greater with large portions than with age-appropriate or self-selected portions. *American Journal of Clinical Nutrition*, *77*, 1164–1170.

French, S. A., Story, M., and Jeffery, R. W. (2001). Environmental influences on eating and physical activity. *Annual Review of Public Health*, *22*, 309–325.

Furst, T., Connors, M., Bisogni, C. A., Sobal, J., and Falk, L.W. (1996). Food choice: A conceptual model of the process. *Appetite*, *26*, 247–266.

Garg, N., Wansink, B., and Inman, J. J. (2007). The influence of incidental affect on consumers' food intake. *Journal of Marketing*, *71*(1), 194–206.

Gregory, R. L. (1972). Cognitive contours. *Nature*, *238*(5358), 51–52.

Herman, C. P., and Polivy, J. (1984). A boundary model for the regulation of eating. In A. J. Stunkard and E. Stellar (Eds.), *Eating and its disorders* (pp. 141–156). New York: Raven.

Hill, A. J., Magson, L. D., and Blundell, J. E. (1984). Hunger and palatability: Tracking ratings of subjective experience before, during and after the consumption of preferred and less preferred food. *Appetite*, *5*(4), 361–371.

International Food Information Council (IFIC). (2008). Promoting health at the Center for Nutrition Policy and Promotion: An interview with Brian Wansink. *Food Insight*, Nov.–Dec., 1, 4–5.

Jansen, A., and van den Hout, M. (1991). On being led into temptation: "Counterregulation" of dieters after smelling a "preload." *Addictive Behaviors*, *16*(5), 247–253.

Kahn, B. E., and Wansink, B. (2004). The influence of assortment structure on perceived variety and consumption quantities. *Journal of Consumer Research*, 30, 581–596.

Klajner, F., Herman, P., Polivy, J., and Chhabra, R. (1981). Human obesity, diet, and anticipatory salivation to food. *Physiology and Behavior*, *27*(2), 195–198.

Langer, E. J. (1990). *Mindfulness*. New York: DeCapo.

Lowe, M. R. (1993). The effects of dieting on eating behavior: A three-factor model. *Psychological Bulletin*, *114*, 100–121.

Meiselman, H. L. (1992). Obstacles to studying real people eating real meals in real situations. *Appetite*, *19*(1), 84–86.

Neisser, U. (1967). *Cognitive Psychology*. East Norwalk, CT: Appleton-Century-Crofts.

Neuendorf, K. A. (2002). *The content analysis guidebook*. Thousand Oaks, CA: Sage.

Nisbett, R. E., and Wilson T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259.

Oppenheimer, D. M. (2004). Spontaneous discounting of availability in frequency judgment tasks. *Psychological Science*, *15*, 100–105.

Painter, J. E., Wansink, B., and Hieggelke, J. B. (2002). How visibility and convenience influence candy consumption. *Appetite*, *38*(3), 237–238.

Pandelaere, M., and Hoorens, V. (2006). The effect of category focus at encoding on category frequency estimation strategies. *Memory and Cognition*, *34*, 28–40.

Polivy, J., and Herman, C. P. (2002). Causes of eating disorders. *Annual Review of Psychology*, *53*, 187–213.

Polivy, J., Herman, C. P., Hackett, R., and Kuleshnyk, I. (1986). The effects of self-attention and public attention on eating in restrained and unrestrained subjects. *Journal of Personality and Social Psychology*, *50*, 1203–1224.

Pronin, E. (2008). How we see ourselves and how we see others. *Science*, *320*, 1177–1180.

Pronin, E., Berger, J., and Molouki, S. (2007). Alone in a crowd of sheep: Asymmetric perceptions of conformity and their roots in an introspection illusion. *Journal of Personality and Social Psychology*, *92*(4), 585–595.

Pronin, E., and Kugler, M. (2007). Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot. *Journal of Experimental Social Psychology*, *43*(4), 565–578.

Rao, A. R., and Monroe, K. B. (1988). The moderating effect of prior knowledge on cue utilization in product evaluations. *Journal of Consumer Research*, *15*(2), 253–264.

Rappaport, L., Peters, G. R,, Downey, R., McCann, T., and Huff-Corzine, L. (1993). Gender and age differences in food cognition. *Appetite*, *20*, 33–52.

Rolls, B. J., Ello-Martin, J., and Ledikv, J. (2005). Portion size and food intake. In D. J. Mela (Ed.), *Food, diet and obesity* (pp. 160–176). Cambridge: Woodhead Publishing.

Rolls, B. J., Engell, D., and Birch, L. L. (2000). Serving portion size influences 5-year-old but not 3-year-old children's food intakes. *Journal of the American Dietetic Association*, *100*, 232–234.

Rolls, B. J., Roe, L. S., Meengs, J. S., and Wall, D. E. (2004). Increasing the portion size of a sandwich increases energy intake. *Journal of the American Dietetic Association*, *104*(3), 367–372.

Rolls, B. J., Rowe, E. A., Rolls, E. T., Kingston, B., Megson, A. and Gunary, R. (1981). Variety in a meal enhances food intake in man. *Physiology and Behavior*, *26*(2), 215–221.

Ross, L., and Nisbett, R. E. (1991). *The person and the situation: Perspectives of social psychology*. New York McGraw-Hill.

Rozin, P., Dow, S., Moscovitch, M., and Rajaram, S. (1998). What causes humans to begin and end a meal? A role for memory for what has been eaten, as evidenced by a study of multiple meal eating in amnesic patients. *Psychological Science*, *9*, 392–396.

Rozin, P., Kabnick, K., Pete, E., Fischler, C., and Shields, C. (2003). The ecology of eating: Smaller portion sizes in France than in the United States help explain the French paradox. *Psychological Science*, *14*, 450–454.

Rozin, P., and Tuorila, H. (1993). Simultaneous and temporal contextual influences on food acceptance. *Food Quality and Preference*, *4*, 11–20.

Schachter, S. (1971). *Emotion, obesity, and crime*. New York: Academic Press.

Schwarz, N. (1996). *Cognition and communication: Judgmental biases, research methods and the logic of conversation*. Mahwah, NJ: Lawrence Erlbaum.

———. (1998). Warmer and more social: Recent developments in cognitive social psychology. *Annual Review of Sociology*, *24*, 239–264.

Sobal, J., and Wansink, B. (2007). Kitchenscapes, tablescapes, platescapes, and foodscapes: Influences of microscale built environments on food intake. *Environment and Behavior*, *39*(1), 124–142.

Squires, S. (2008, February 12). Bringing nutrition home. *Washington Post*, pp. F-1+.

Stroebele, N., and De Castro, J. M. (2004). Effect of ambience on food intake and food choice. *Nutrition*, *20*, 821–838.

Terry, K., and Beck, S. (1985). Eating style and food storage habits in the home: Assessment of obese and nonobese families. *Behavior Modification*, *9*(2), 242–261.

Vartanian, L. R., Herman, C. P., and Wansink, B. (2008). Are we aware of the external factors that influence our food intake? *Health Psychology*, *27*(5), 533–538.

Volkow, N., Wang, G., Fowler, J. S., Logan, J., Jayne, M., Franceschi, D., Wong, C., Gatley, S. J., Gifford, A. N., Ding, Y., and Pappas, N. (2002). "Nonhedonic" food motivation in humans involves dopamine in the dorsal striatum and methylphenidate amplifies this effect. *Synapse*, *44*(3), 175–180.

Wansink, B. (1994). Antecedents and mediators of eating bouts. *Family and Consumer Sciences Research Journal*, *23*, (2), 166–182.

———. (1996). Can package size accelerate usage volume? *Journal of Marketing, 60*(3), 1–14.

———. (2004). Environmental factors that increase the food intake and consumption volume of unknowing consumers. *Annual Review of Nutrition*, *24*, 455–479.

———. (2006a), *Mindless eating: Why we eat more than we think*. New York: Bantam-Dell.

———. (2006b). Nutritional gatekeepers and the 72% solution. *Journal of the American Dietetic Association*, *106*(9), 1324–1326.

———. (2008). Project M.O.M.: Mothers & others & MyPyramid. *Journal of the American Dietetic Association*, *108*(8), 1302–1304.

———. (2009, January 16). Until we meet again. . . . *MyPyramid e-Post*. Retrieved from http://www.docstoc.com/docs/71892015/MyPyramid-e-Post

Wansink, B., and Cheney M. M. (2005). Super bowls: Serving bowl size and food consumption. *Journal of the American Medical Association*, *293*(14), 1727–1728.

Wansink, B., and Deshpande, R. (1994). "Out of sight, out of mind": Pantry stockpiling and brand-usage frequency. *Marketing Letters*, *5*(1), 91–100.

Wansink, B., Just, D., and Payne, C. (2009). Mindless eating and healthy heuristics for the irrational. *American Economic Review*, *99*, 165–169.

———. (2010). *Constrained volition and everyday decisions.* Manuscript submitted for publication.

Wansink, B., and Kim, J. (2005). Bad popcorn in big buckets: Portion size can influence intake as much as taste. *Journal of Nutrition Education and Behavior*, *37*(5), 242–245.

Wansink, B., Painter, J. E., and Lee, Y. K. (2006). The office candy dish: Proximity's influence on estimated and actual consumption. *International Journal of Obesity*, *30*, 871–875.

Wansink, B., Painter, J. E., and North, J. (2005). Bottomless bowls: Why visual cues of portion size may influence intake. *Obesity Research*, *13*(1), 93–100.

Wansink, B., and Park, S. B. (2001). At the movies: How external cues and perceived taste impact consumption volume. *Food Quality and Preference*, *12*(1), 69–74.

Wansink, B., Payne, C. R., and Chandon, P. (2007). Internal and external cues of meal cessation: The French paradox redux? *Obesity*, *15*, 2920–2924.

Wansink, B., and Sobal, J. (2007). Mindless eating: The 200 daily food decisions we overlook. *Environment and Behavior*, *39*(1), 106–123.

Wansink, B., and van Ittersum, K. (2003). Bottoms up! The influence of elongation and pouring on consumption volume. *Journal of Consumer Research*, *30*(3), 455–463.

Wansink, B., and van Ittersum, K. (2007). Portion size me: Downsizing our consumption norms. *Journal of the American Dietetic Association*, *107*(7), 1103–1106.

Wansink, B., and van Ittersum, K. (2005). Shape of glass and amount of alcohol poured: Comparative study of effect of practice and concentration. *British Medical Journal*, *331*(7531), 1512–1514.

Ward, A., and Mann, T. (2000). Don't mind if I do: Disinhibited eating under cognitive load. *Journal of Personality and Social Psychology*, *78*, 753–763.

Webber, R. P. (1989). *Basic content analysis.* Newbury Park, CA: Sage.

World Health Organization (WHO). (1998). *Obesity: Preventing and managing a global epidemic*. Geneva: World Health Organization.

Young, L. R. (2000). *Portion sizes in the American food supply: Issues and implications* (Doctoral dissertation). New York University.

Young, L. R., and Nestle, M. (2002). The contribution of expanding portion sizes to the US obesity epidemic. *American Journal of Public Health, 92*(2), 246–249.

CHAPTER 19

A Social Psychological Approach to Educational Intervention

JULIO GARCIA

GEOFFREY L. COHEN

The causes of academic underperformance are a major concern of the educational community and policy makers in the United States. Of particular importance is the achievement gap between at-risk minority students and European American students and its potential remedies. Academically at-risk minority students, such as African Americans and Latino Americans, perform almost a standard deviation below European American students on intelligence tests and earn school grades below those of their European American peers (Jencks and Phillips, 1998; Nisbett, 2009). Between the years 2004 and 2007, while 6 out of every 100 European American young adults had not received a high school diploma or its equivalent, the corresponding figures for African Americans and Latino Americans were, respectively, 10 and 22 out of every 100 in their ethnic group (U.S. Department of Education, 2009). These achievement gaps persist in spite of the local and national initiatives aimed at closing them (Dillon, 2006; Neal, 2005). In a society such as that of the United States, where economic opportunity depends heavily on scholastic success, even a partial remediation of the achievement gap would lead to a positive change in the lives of many at-risk children.

Our research focuses on the impact that social-psychological factors have on the academic outcomes reflected in this gap (Cohen and Garcia, 2008; Cohen et al., 2006; Walton and Cohen, 2007). At the heart of our effort lies the notion of the classroom as a tension system in which various factors, including structural factors such as classroom size and psychological factors such as student perceptions, interact to produce a stable environment that elicits a consistent set of attitudes, behaviors, and outcomes over time. Differences among groups arise from consistent differences both in their objective experience and in their subjective perceptions. These notions can aid us in understanding student performance and help us to develop effective educational practices. They suggest an approach to the achievement gap that we have found to be productive and that has important social policy implications for addressing the pressing social problem of underperformance. Before reviewing our work, we will discuss the idea of a tension system and its relevance to academic performance.

The Classroom as a Social Tension System

Social environments, classrooms included, can be viewed as tension systems consisting of forces in a dynamic state of interaction that remains relatively stable over time (Lewin, 1948, 1951; Ross and Nisbett, 1991). Generally social tension systems transcend single instances. Children, for example, expect to be in a classroom with their teacher through the year. Also, tension systems consist of forces that are unique to them and of other forces that are more general in nature, such as cultural norms and moral codes. In the United States, for instance, the classroom is seen as an environment designed to develop the appropriate and necessary social and intellectual competencies of individuals of a particular developmental stage. It is assumed that a number of forces or factors will be present to promote this goal, including trained instructors, appropriate teaching materials, an adequate physical space, and a program of learning that consists of goals and milestones. Beyond these general factors are others unique to individual classrooms, such as the teacher's personality, the demographic makeup of the students in the classroom, the curriculum priorities of the school, and the administrative leadership.

The forces in a social tension system can facilitate or restrain a given outcome. In both the classroom and the larger school environment, there are a number of

forces that can help or hinder academic performance. Schools are systems that, while designed to promote learning, can also contain forces that make their tasks difficult, or in some cases impossible, to accomplish. For example, a school can lack sufficient material resources to provide students and teachers the necessary tools to reach the desired level of performance, or the environment may be so threatening that students and teachers are unable to overcome it. In interaction, these forces shape the learning environment and determine the overall level of performance of its students. At a macro level, overarching social factors in the form of educational policy, social organization, and political ideology can constitute facilitating or restraining forces in the classroom. Ideas about appropriate class size, the importance of standardized testing, and the impact of socioeconomic class, gender, and ethnic distinctions affect the classroom environment to one degree or another. For instance, the accountability movement, which issued from theories of academic achievement and achievement gaps, has had a large impact on the classroom through its effect on curriculum priorities, teaching methods, and the frequency of standardized testing.

A social tension system that appears fixed has reached a point at which the interaction between its facilitating and inhibiting forces has stabilized. However, this state of balance can be altered or tipped by any number of events that trigger a change in the relationship among the factors, thus leading to a new state of balance or status quo. For instance, the intensity of a particular force can change, or a new force can be introduced. Student motivation could increase, for example, when an esteemed role model visits the school (Lockwood and Kunda, 1997), or teacher preparation could be raised with an increase in subsidies for professional development. The point is that while a social environment, like a classroom, may appear relatively static and resistant to change, it often is not.

To bring about change in an environment, three aspects of tension systems need to be kept in mind. First, because tension systems involve complex interactions among forces, individual forces can impede or amplify one another (Ross and Nisbett, 1991). Social approval from peers, for example, can facilitate school achievement in a context where such approval is tied to academic success. On the other hand, social approval can restrain school achievement in contexts where peers disapprove of academic success or where their approval can be more readily won in some other domain, such as sports. The effect of a given force, in this case social approval, is in large part dependent on context. For instance, Fryer and Torelli (2005) found that academic success was associated with lower popularity for ethnic minority students when they attended predominately White schools, but not when they were in predominately minority urban schools. Although many interpretations are plausible, differences in the salience of race in these different types of schools may affect how high achievement is perceived by students.

Because of the interactive nature of tension systems, processes can, for good or ill, feed off one another's effects. This can convert small initial differences between individuals and groups into large and long-term ones, thus exacerbating inequality. This is especially the case in environments that allocate rewards and punishments based on merit and that define merit largely in terms of observable performance along a few set criteria. For instance, students who begin school slightly ahead in academic preparation may be given opportunities and provided with higher expectations, while their low-achieving peers are assigned to low-expectation tracks and viewed as less able and less worthy of attention and mentoring (Rosenthal and Jacobson, 1992; see also Jussim and Harber, 2005; Woodhead, 1988). As a consequence, lower-achieving students could then perform still worse, which in turn could reinforce teachers' expectations, in a potentially repeating cycle. Each of these situations reflects a recursive process in which the "rich get richer" or "poor get poorer." In this way, an "underachieving environment" can emerge in the latter case and call forth consistent underperformance from some groups of students. Some support for this notion is found in the finding that low-achieving boys entering a new grade may show a large gain in performance by apparently leaving behind the norms, expectancies, and channels of their previous classroom (see Dweck et al., 1978). This cycle may also help to explain the downward spiral in performance commonly observed in junior high school (Eccles, Lord, and Midgley, 1991), particularly among minority students (Simmons, Black, and Zhou, 1991). This period is a time when school becomes more evaluative, the performance standards shift upward, and failure processes become more likely to feed off of one another.

The second aspect of tension systems to keep in mind when attempting to effect change is that many of the forces in a system go unobserved or underappreciated until efforts to change it are made (Ross and Nisbett, 1991). As Kurt Lewin remarked, "If you want truly to understand something, try to change it." The *Move to Opportunity* program provides an example of this phenomenon. The program was designed in part to offer disadvantaged children educational opportunities by providing poor families the chance to move to less impoverished neighborhoods. This program has had many positive effects, but the

hoped-for long-term effects on children's academic test scores did not materialize (e.g., see Sanbonmatsu et al., 2006). This lack of improvement may have occurred because of underappreciated restraining forces involved in the situation. On moving to their new neighborhoods, poor families are faced with a number of pressing priorities, such as remaining close to relatives and friends, which can restrain their ability to identify and act on the new academic opportunities available to their children.

Finally, although objective structural factors obviously affect behavior, the mental and psychological processes of individuals are also critical elements in a social tension system and thus must be considered in predicting the effects of such systems on behavior (Ross and Nisbett, 1991). With respect to the classroom, while such processes include the student's level of intellectual ability, psychological factors not directly related to ability can also affect performance. These form what we call the individual's *psychological environment*—that is, their perceptions of themselves and their environment. Among the most important of these are factors related to people's perceptions of the fairness of their social environment (Tyler, this volume; see also Cohen and Steele, 2002; Huo et al., 1996; Tyler, 2004). Indeed, perceptions of whether fair procedures are used in making decisions and allocating rewards and punishments are consistently a better predictor of compliance and internalization of organizational norms than are the actual allocated rewards and punishments (Huo et al., 1996). Also, one of the strongest predictors of people's compliance with authorities in an organization, such as students' compliance with teachers in their school, is their perception of procedural justice, the perceived fairness of the processes and procedures in their environment.

Moreover, the social psychologists Al Bandura and Carol Dweck have documented how psychological processes can shape students' perceptions of the academic environment and affect their intellectual performance (Bandura, 1986; Dweck, 1999). For instance, two children with the same level of ability and confronted with the identical objective level of failure can respond in a completely different ways due to differences in their psychological functioning. Students with low self-efficacy—those who doubt their ability to succeed in school—or students who believe that their level of intelligence is a fixed quality, are more likely than their peers to give up, persevere in ineffective strategies, experience negative emotion, and fail to return to their original performance level following failure. By contrast, students with high self-efficacy, or those who believe that intelligence is a malleable quality that expands with practice, are more likely to view a situation as a challenge, try harder, entertain novel strategies, and return to and even exceed their original performance level.

In summary, both social structural factors and psychological factors have a large impact on performance. Many psychological factors, as we will see, can act as powerful restraining forces, preventing positive forces in both the student and the environment from asserting their full impact on behavior. Just as drag can prevent a car from achieving its top speed and efficiency, psychological forces can lessen the efficacy of a school system. Psychological forces can, on the other hand, also have substantial impact by acting as tipping or triggering agents that permit the positive forces to fully assert themselves.

The Minority Achievement Gap

The notion of the school as a dynamic tension system informs many current educational initiatives, interventions, and policy aims. It is evident in policies to reduce the number of students in a class, provide school meals, and increase parents' involvement in their children's education. These policies assume that the school environment is complex, and that key environmental factors interacting with the student affect the system's overall performance.

The notion of school as a tension system is also evident in analyses of the persistent achievement gaps found in American classrooms. One of the more accepted explanations for the gap in academic achievement between White and Asian students on the one hand, and their African American and Latino American peers on the other, is that it is primarily due to differences in socioeconomic status (SES). Central to this explanation is the idea that there are factors linked to SES that can interact with the classroom in ways that affect a child's academic performance. Among these are the presence of college-educated adults who can serve as role models or resources, the availability of books in the home, the level of vocabulary and the amount of social engagement, Socratic questioning, and negotiation that occurs in the family (Brooks-Gunn and Furstenberg, 1986; Gordon and Lemons, 1997; Hart and Risley, 1995). While low SES does predict lower academic performance, it does not sufficiently explain the performance differences between certain groups. Critically the SES explanation offers a testable hypothesis that can be stated as follows: when a significant number of individuals from these lower performing racial or ethnic groups attain middle-class SES and above, the performance differences between them and European-Americans and Asian-Americans will diminish significantly or cease to exist. Much to the disappointment of many, the authors included,

this has not occurred to the degree one would expect given changes in the economic status of racial and ethnic minorities. At every level of social economic status in the United States, the racial and ethnic achievement gap persists in spite of the increasing number of minority individuals attaining middle-class and higher status levels (Hacker, 1995; Jencks and Phillips, 1998; Nisbett, 2009; Steele, 1997; see also Bowen and Bok, 1998).

Given this, we revisited the problem of the achievement gap to reconsider the factors at work in the classroom and how these might interact to produce the gap. Our thinking shares the emphasis on the importance of the situation at the heart of the SES explanation: the individuals in lower-performing ethnic and racial groups are not inherently less capable of performing well.

A Social Psychological Constraint on Performance: Identity Threat

The work of Claude Steele and his colleagues provided an intellectual underpinning for our initial thinking and the research results to buttress it. In a series of what have become seminal studies, Steele and his associates Joshua Aronson and Steve Spencer demonstrated that the achievement gap between African Americans and their European American peers on standardized intellectual tests, and between males and females on the math portion of these tests, could be dramatically lessened by altering the *psychological environment* (Steele, Spencer, and Aronson, 2002; see also Davies, Spencer, and Steele, 2005; Schmader, Johns, and Forbes, 2008).

Members of such groups may worry that their poor performance could confirm the negative stereotype about their group in the eyes of others, a preoccupation called *stereotype threat* (Steele, Spencer, and Aronson, 2002). This threat can cause stress that undermines performance. As a consequence, altering the psychological environment to render the stereotype irrelevant can boost performance. In a study conducted by Steele and Aronson (1995), African American college students were told that the Graduate Record Exam (GRE) they were about to take was "diagnostic of academic ability." This raised the possibility for them that they could reinforce a negative stereotype about their race's intelligence if they performed poorly. This preoccupation led African Americans students to perform at only half the level of European American students, controlling for prior ability level as roughly measured by previous test scores. However, African Americans' performance equaled that of European American students (again controlling for prior ability level) when the same test was presented as "non-diagnostic of ability," that is, irrelevant to the stereotype. Similar effects were shown for the performance of female college students on a difficult standardized math test in a series of studies conducted by Spencer, Steele, and Quinn (1999). Women's performance on a math test was significantly lower than that of their male peers. By contrast, when informed that the same test produced no gender differences—that men and women performed equally on it—women achieved a level of performance equal to that of men. Such effects have been documented among other stereotyped groups, including Latino Americans (Schmader and Johns, 2003; see also Aronson, 2002) and low-SES students in school (Croizet and Claire, 1998), high-performing White students reminded of the stereotype of Asian superiority in math (Aronson et al., 1999), and White men in the domain of sports (Stone et al., 1999). Stereotype threat has been replicated in more than a hundred studies and tends to occur on relatively difficult tasks that pose the risk of confirming a stereotype (Ben-Zeev, Fein, and Inzlicht, 2004; O'Brien and Crandall, 2003; Spencer, Steele, and Quinn, 1999). Among the replications are recent studies by a variety of investigators (e.g., Grimm et al., 2009; Rydell, McConnell, and Beilock, 2009; for reviews, see Schmader, Johns, and Forbes, 2008; Shapiro and Neuberg, 2007; Steele, Spencer, and Aronson, 2002; Walton and Cohen, 2007; Walton and Spencer, 2009).

This research provided a basis for our examination of the classroom as a social tension system. It highlights the idea that if outcomes differ systematically for groups of individuals in a social environment, then what appears to be the same environment for everyone may in fact be different. That is, social environments can differ radically both objectively and psychologically for the groups in them. It is not difficult to think of ways that this could be true in a classroom for individuals of certain racial or ethnic groups. At the objective level it is possible, due to discrimination, that such individuals could receive fewer material resources, be given less access to teachers or other learning specialists, or be held to lower standards than their White peers.

However, even in classrooms where the environment does not differ in any apparently objective way, the psychological or subjective environment can differ for individuals in these groups. The awareness that racial prejudices might be in play could make for a different psychological environment for stereotyped students. To begin with, it would be an environment where their group or social identity would be, for better or worse, salient to them (Cohen and Garcia, 2005; Steele and Aronson, 1995). This salience could

call forth a host of attitudes and behaviors associated with that identity, including a sense of solidarity and a set of coping behaviors. It could also give rise to chronic concerns not only that they may be judged in light of a negative stereotype about their group, but also that fellow group members may be so judged as well—a preoccupation termed *collective threat* (Steele, 1997; Steele and Aronson, 1995; Steele, Spencer, and Aronson, 2002; see also Aronson, 2002; Aronson and Inzlicht, 2004; Cohen and Garcia, 2005; Cohen and Steele, 2002; Cohen, Steele, and Ross, 1999). Such concerns can arise irrespective of the actual level of prejudice and discrimination in an environment.

For racial and ethnic minorities who find themselves the target of negative stereotypes that place their intellectual abilities under suspicion, the psychological environment of the classroom is one in which their identity is at risk in at least two ways. First, it can be threatening to their self-worth, regardless of their race or ethnicity, because of the constant evaluation of their skills and the specter of possible poor performance and its consequences. We are not suggesting that such evaluation is necessarily bad, only that it can be stressful. Second, the environment can also threaten them by raising the possibility that a valued aspect of their identity, their group, will be devalued. This is something White students do not generally experience in the classroom. Because there are two sources of stress for minority students, the normal stress associated with a chronically evaluative situation and the stress linked to their social identity, it is more likely that these students could reach stress levels that inhibit their performance. Interestingly, those who are highly identified with academics and invested in doing well are often the most likely to suffer such performance-inhibiting anxieties (Marx, Brown, and Steele, 1999; Steele, 1997).

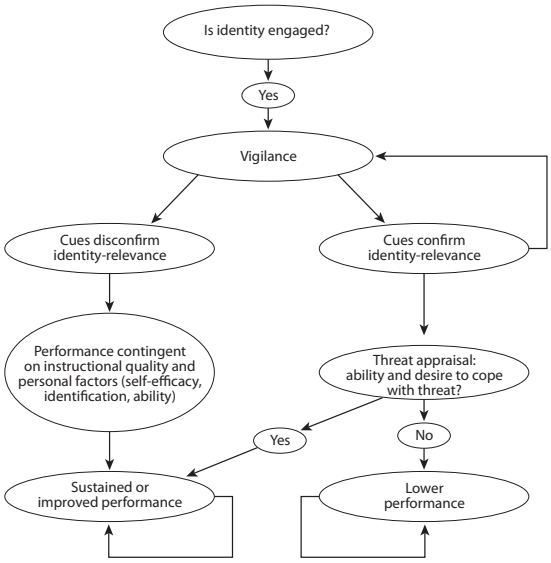
There are other aspects of the classroom that can be particularly troubling for minority individuals that are not generally present in the classroom environment of Whites (see also Branscombe, Schmitt, and Harvey, 1999). These students cannot necessarily lessen the threat to their identity by a strong performance, since they may understand that those holding a negative stereotype will often discount counterstereotypic behavior. These others may characterize those who perform well as exceptions to the rule (Richards and Hewstone, 2001) or single out the behavior of a single minority that confirms the stereotype (Henderson-King and Nisbett, 1996). Such knowledge can lessen the likelihood that they, in spite of having performed well, will benefit from a positive recursive cycle in which high performance sustains itself or promotes even higher performance. Also, these students understand that regardless of how well they do, there will always be some individuals in their group who will perform poorly and potentially provide evidence in support of the denigrating stereotype. We have shown that the mere possibility that a fellow group member could do poorly on an intellectual task can change the psychological environment for an individual and trigger performance-debilitating psychological effects even among elite college students (Cohen and Garcia, 2005).

Not only do the social environments of at-risk minority students and White students differ in critical ways, but the environment in which the former students function makes it more likely that they will suffer long-term performance deficits. Like their White peers, their skills as students are continually being evaluated, and so they are subject to all the psychological consequences that follow from being in such an environment. However, beyond the possible aversive consequences present for them personally, they carry an additional burden. They must also contend with the potentially aversive consequences that the environment holds for their group, and by extension, that aspect of their identity related to their group, known as their social identity. Because their group is the target of a negative stereotype regarding the intellectual abilities of its members, these students must be concerned about whether they and their fellow group members will be judged in light of this widely known negative social judgment. This can then intensify psychological factors such as stress that inhibit motivation and performance. It will also increase the chance that poor performance will yield still poorer performance in a prolonged recursive process.

The Identity Engagement Process

The presence of an *identity threat* for targeted minority students that their White peers do not experience underlies differences in the psychological environments between these groups and is key to understanding the differences in their performance. While it is imperative to gain a greater understanding of identity-threat processes and how these interact with other factors in a range of social environments, the urgency and importance of the issue of minority achievement leads us to focus on examining these processes in school (see Cohen et al., 2007, for the role of similar processes in intergroup conflict).

Obviously there are a number of factors that identity threat could interact with in the classroom. However, we will limit our examination here to those factors that could interact with identity threat in a way that affects the academic performance of minority students. Our general notions about how such



19.1. The identity engagement model.

interactions can play out in the social context of the classroom are introduced in figure 19.1 (see Cohen and Garcia, 2008). It displays our model of how identity processes can affect performance and is more fully developed in the discussion that follows.

Upon entering an important social environment, like a classroom, an individual tends to make a general assessment. He or she asks, "Is this a situation in which my identity could be a factor in my outcomes?" If the answer is yes, the person's identity will be psychologically engaged. Here we focus on cases where the person's identity has possibly negative, rather than positive, consequences for the individual. For instance, most African Americans know that school and work fall into a class of situations in which they could be judged negatively because of their race, whereas certain sports settings are situations in which they could be seen positively because of their race (Steele and Aronson, 1995; Walton and Cohen, 2007). Assessments of the environment are often made nonconsciously and informed by personal experience, historical knowledge, and socialization.

People tend to become vigilant in environments where their identity is engaged (Frable, Blackstone, and Scherbaum, 1990; Kaiser, Brooke, and Major, 2006; Purdie-Vaughns et al., 2008). They monitor such situations for cues related to whether their identity is relevant to their outcomes, for instance, whether it affects how they are treated by important figures in their social environment. A minority student, for example, might scrutinize a teacher's nonverbal behavior or feedback for evidence of bias (Cohen and Steele, 2002; Crocker and Major, 1989). As in any hypothesis-testing process, people may be more sensitive to bias-confirming evidence than to bias-disconfirming evidence (Darley and Gross, 1983; Kleck and Strenta, 1980; Walton and Cohen, 2007). Vigilance of this sort is a general and adaptive process. If one believes that one could be treated poorly or unfairly, it is adaptive to monitor for the possibility of such treatment until one either is given or gathers information to the contrary. For instance, a person might suspect that a superior doubts his or her abilities and could disparage or penalize him or her openly or behind closed doors. In encounters with this superior, it would be both natural and adaptive for the individual to attend to whether this situation is one in which the superior's suspected predisposition will play a role. The individual might focus on the formality of the superior's greeting, his or her nonverbal behavior, such as body language, the valence or quality of the superior's feedback, or how the superior treats others. This would enable the individual to prepare, both psychologically and behaviorally, for an aversive situation.

If the cues *disconfirm* the relevance of their identity to the situation, people will tend to feel that they are viewed as individuals, and their performance will depend on structural and personal factors such as the quality of instruction and their level of self-efficacy, identification with school, and skill. In one study, for instance, African American students responded as positively as White students to critical feedback when it was made clear that the critical nature of the feedback was motivated not by racial bias but by high standards and a belief in their ability to *reach* those standards (Cohen, Steele, and Ross, 1999). When the threat of group-based devaluation was disconfirmed, students could better avail themselves of the learning opportunities in the feedback.

If, on the other hand, the cues *confirm* the relevance of their identity to the situation, a threat-appraisal phase follows. People will assess whether they have the ability to deal with the threat, and, if they do, whether they want to do so (Lazarus and Cohen, 1977). Students might see the degree of bias in a classroom as surpassing their ability or desire to overcome it. If so, their performance will suffer either directly by lowering the motivation to perform or indirectly by triggering psychological factors, such as stress, that undermine performance. One possible outcome is *disidentification*, or *devaluation*, in which students downplay the importance of school or of the criteria being used to evaluate their merit (Schmader, Major, and Gramzow, 2001; Steele, 1997). If, on the other hand, students perceive that they have the ability and desire to contend with the threat, this could

lead to maintained or improved performance, as individuals marshal the psychological resources to meet the challenge (Cohen and Garcia, 2005).

A key aspect of this process is *recursion*, a cycle of repeated steps or outcomes, each based on the result of the one before. By definition, the consequences of such a cycle can themselves become causes for subsequent behavior. In the case of the classroom, a student's performance could not only directly affect him or her, but also, because it is socially interpreted and acted upon, lead to feedback from others that then further affects the person's performance. For instance, a chronically underperforming student may be viewed by teachers as less able or may be assigned to a lower academic track, either of which could inhibit later performance (Rosenthal and Jacobson, 1992). Negative recursion tends to occur in chronically evaluative settings in which opportunities are allocated based on a small set of behavioral criteria used to assess merit. Recursion, with similar outcomes following one another, occurs not only at the social level but also at the psychological level. When students perform well, they feel efficacious and less threatened, and as a result they perform better the next time, which in its turn can make them feel more efficacious (Bandura, 1986). Likewise, when students perform poorly, for example as a result of stress, even worse performance can follow due to increases in stress, threat, and vigilance (Wilson, Damiani, and Shelton, 2002).

The recursive nature of chronically evaluative environments also offers an opportunity. Because a recursive process depends on a continuous feedback loop, an early interruption of that loop could produce long-term benefits. Additionally, because even small early gains in performance compound with time, the recursive process can be turned to one's advantage, increasing the likelihood and longevity of student success (Cohen et al., 2006, 2009).

Interventions

The process of identity engagement suggests four important approaches to intervention. The first, most obviously, involves reducing prejudice and stereotyping, so as to change the actual and perceived outcomes associated with a social identity. This is done in programs designed to reduce prejudice in schools, such as the jigsaw classroom described later in this chapter (Aronson and Patnoe, 1997). The second approach highlights the value of changing the failure process in the social environment, so as to block downward recursive processes fueled by the social environment (Woodhead, 1988). Such changes could include substituting remedial programs with programs that challenge even low-performing students with high standards (Fullilove and Treisman, 1990; Steele, 1997; Steele et al., 2004). Modifications could also involve broadening the criteria of merit by using alternative modes of assessment, for instance student portfolios, that are less susceptible to identity-threat processes (see Tierney et al., 2003).

Reducing an individual's tendency to interpret their experience in light of social identity at the vigilance stage and buffering individuals from any detrimental psychological or emotional impact of this tendency at the threat-appraisal stage are two psychological intervention strategies suggested by the identity engagement process (Cohen and Garcia, 2008). Contrary to common wisdom, neither approach involves directly confronting the stereotype. Indeed, it is possible that doing so may do more harm than good, because directly raising the stereotype may be distressing to some individuals. Below we will review research featuring randomized, double-blind experimental designs testing these two strategies in real-world classrooms.

Taking Social Identity off the Table at the Vigilance Stage

This strategy involves helping students to make constructive attributions for the challenges they face in school through the use of *attributional retraining* (see Wilson et al., 2002; Wilson and Linville, 1985). Students are taught to attribute adversity and hardship to factors not directly relevant to race, the stereotype, or a personal lack of ability or sense of belonging. Instead they are encouraged to attribute adversity and hardship to the challenges inherent in school. In one of the experimental conditions in a study by Good, Aronson, and Inzlicht (2003), for example, students were exposed to role models who discussed their initial difficulties after moving from elementary to middle school but who reported getting increasingly better grades as they learned the ropes and kept working. In another experimental condition, they were led to view intelligence as expandable rather than fixed, lessening the tendency to see frustration in school as evidence of intellectual limitation (see also Aronson, Fried, and Good, 2002; Blackwell, Trzesniewski, and Dweck, 2007). Compared to students in a control group, students in both conditions went on to earn higher statewide test scores. Similar positive effects of such interventions on grades were displayed in a New York City school by low-achieving African and Latino American students from economically disadvantaged backgrounds (Blackwell, Trzesniewski, and Dweck, 2007).

In another experiment, freshmen at a predominately White university were asked, at the end of the

difficult first year of college, to review the results of a survey of upperclassmen at their school (Walton and Cohen, 2007). The results conveyed, first, that almost all students regardless of race felt uncertain of their belonging in the first year of college and, second, that these doubts lessened with time. The results led students to view their doubts about belonging as a common occurrence rather than unique to them or members of their racial group, and as transitory rather than fixed. For instance, as one student stated in the survey, "I worried that I was different from other students. . . . Now it seems ironic—everybody feels they are different freshman year from everybody else, when really in at least some ways we are all pretty similar." Students were also led to internalize the message by giving a speech in front of a video camera, ostensibly for viewing by incoming freshmen, describing how their experiences were consistent with these survey results. Students thus came to see the difficulties they were experiencing in school as part of the normal learning curve that most students go through when they enter a new environment and face new challenges. While the intervention had no consistent effect on Whites, it buttressed African Americans' sense of belonging on days of hardship. Additionally, in the following semester, intervention-treated African Americans earned a higher college GPA, an effect that follow-up data indicate persisted into students' junior year.

Lessening the Impact of Social Identity Threat at the Appraisal Stage

Instead of affecting people's sensitivity to the possibility of being stereotyped, the second psychological strategy demonstrates the efficacy of intervening at the threat-appraisal stage by increasing people's psychological resources. Underpinning this strategy is the notion that people want and need to see themselves in a positive light—to have a sense of self-integrity. In other words, people want to believe that they are good people and that they can cope with their environments. Moreover, it is possible to assure people that they do indeed have self-integrity by having them engage in *self-affirmations*. In this process people reinforce self-integrity by reflecting on important domains of identity unrelated to the provoking stressor (Steele, 1988; see also Sherman and Cohen, 2006). People are better able to cope with threat in one domain, school for instance, if they can shore up their self-integrity in another, such as family. More important, as self-affirmation reduces stress arising from evaluative performance settings (Creswell et al., 2005), we assumed that this, in turn, could improve performance (Martens et al., 2006).

Two field experiments were conducted in a suburban middle-class middle school where African Americans made up approximately 50% of the student body. Seventh-grade students completed an affirmation exercise in class early in the school year, a stressful time. They wrote about a personally important value, such as religion or relationships with friends (Cohen et al., 2006). The exercises, which were usually given before a test or exam, had students integrate the value into their lives in the context of a series of structured writing assignments. Students' writing touched on diverse issues of personal significance. For instance, one student wrote, "[Art] is important to me because it makes me feel calm. When I'm very upset, like I'm going to cry I sit down and start listening to music or start drawing a picture." Another wrote, "My friends and family are most important to me when I have a difficult situation that needs to be talked about. My friends give me companionship and courage. My family gives me love and understanding."

African Americans who had been given the opportunity to self-affirm earned a higher course GPA than students of their race completing control exercises requiring them to write about neutral topics (Cohen et al., 2006). The intervention was associated with a roughly 40% reduction in the race gap in GPA in the course in the fall term. Follow-up data indicate that the intervention had an effect on overall GPA that persisted for at least two years, roughly eliminating 30% of the difference in GPA that had existed between African Americans and European Americans in previous years (Cohen et al., 2009). Perhaps more tellingly, at the practical level the intervention reduced the percentage of African Americans earning a D or below in the first term of the course from 20% to 9%. The latter rate was no different from the rate observed for White students. The potential importance of the latter finding is underscored by the fact that the poorest-performing students in school often require a disproportionate amount of a school system's resources to provide for their needs. Additionally, preliminary follow-up data related to state achievement-test performance indicated that the intervention again benefited African American students' performance. Unlike most other interventions, this intervention most benefited the most "at risk" students, reducing group-based differences in performance while not adversely affecting other students (compare Ceci and Papierno, 2005).

Although these results seem unique, in fact they are not unprecedented. Social psychological research provides ample evidence that seemingly small interventions can have large and long-term effects (Dholakia and Morwitz, 2002; Freedman, 1965; Wilson, 2006;

see also Benartzi, Peleg, and Thaler, this volume; Thaler and Sunstein, 2008). Insofar as psychological interventions appear to have disproportionate impacts in relation to the time, effort, or resources they require, preexisting environmental processes must be instrumental in the transformation of their initial effects into larger and long-term outcomes (Woodhead, 1988).

With regard to the results of our research, one such process involves recursive performance cycles. For instance, as with other effective psychological interventions, the affirmation intervention interrupted a downward performance trajectory (Blackwell et al., 2007; Wilson et al., 2002). The intervention buffered students against the negative consequences of early poor performance, consequences that would otherwise compound into increasingly worse performance as the result of a recursive cycle. The GPAs of minority students in the control group declined throughout seventh and eighth grade, something not uncommon in the middle school years (Eccles, Lord, and Midgeley, 1991). Indeed, the greater the decline in grades prior to the experimental manipulation, the greater the decline later (Cohen et al., 2006). By contrast, the GPAs of intervention-treated African Americans declined less over the two years that were examined. In fact, not long after the first intervention, their grades improved, so that any decline in performance they had experienced prior to the intervention bore no relationship to their later performance. The intervention thus seemed to interrupt a downward trajectory and perhaps initiated another, now positive, recursive cycle.

Even if the effects on performance are initially small, they can become large if they accumulate in an additive fashion across multiple trials or tests. As an analogy, in professional baseball, small differences in the number of successful at-bats during individual games can compound over an entire season and career and lead to one being considered an all-star rather than just another good player (Abelson, 1985). Similarly, in the classroom, a small but consistent intervention effect on individual evaluations can compound into a meaningful effect on final grades.

One way in which relatively small initial performance benefits can be carried forward is through social-psychological processes. Students could, for instance, feel self-affirmed by performing well relative to their standards, even if the improvement was objectively relatively minor, such as going from their usual C– to a C on an exam. As a consequence of being affirmed, the factors inhibiting their performance could be reduced. This could be especially powerful if a trend of increasingly poor performance is interrupted and deflected upward, as was the case with students completing the affirmation intervention (Cohen et al., 2006). Students could see this as particularly strong evidence of their competence and integrity. This reinforcement of their self-efficacy and self-integrity would increase the likelihood that they would at least begin to perform up to their actual skill level.

A meaningful portion of the achievement gap, in our view, is due to social-psychological processes that inhibit minority students from manifesting their actual academic skills. There is some evidence to support this notion in our work. Not long after receiving the affirmation, for the first time since the beginning of the school year, minority students did not experience a decline in their performance. In fact these students displayed nearly the same level of performance as their White peers (Cohen et al., 2006). Social comparison processes may have come into play at this point. The sense of efficacy and integrity of minority students receiving the intervention may be reinforced because they see themselves performing almost as well as their White peers. These students would also have first-hand evidence that intellectual performance is malleable rather than fixed—improvable with effort and practice—a notion that would further their motivation and performance (Dweck, 1999; see also Aronson, Fried, and Good, 2002).

Because they are performing better, these minority students may also become less vigilant in regard to the stereotype and so less likely to interpret their classroom experience in light of it. This would reduce the likelihood that they would experience stereotype threat and the stress associated with it. Consistent with this expectation, the intervention reduced the cognitive accessibility of the racial stereotype among minority students (Cohen et al., 2006). Because the psychological availability of mental concepts affects the encoding of social experience (Fiske and Taylor, 1991), this could in turn have led students to see less bias in their school. Indeed, follow-up data suggest that minority students receiving the intervention proved relatively more likely to maintain their trust in their teachers over the course of the year than did their fellow students (see also Sherman and Cohen, 2006). These perceptions may then have led students to interpret their teacher's behavior more charitably and may have helped to sustain their sense of adequacy in school even in the face of adversity (Cohen et al., 2009; Huo et al., 1996; Tyler, 2004). In summary, as a result of the intervention, the students have gone from an environment in which they could expect only deteriorating performance to an environment in which it is possible to *do well*, and, perhaps more

important, one in which they believe their teachers will recognize their success.

Social processes can also act as factors that facilitate the transformation of initial benefits into long-term ones. Students receiving the intervention, upon performing better, may be seen by their teacher as more able. Such students may then receive more attention, mentoring, and challenge in the classroom (Rosenthal and Jacobson, 1992). They may also be more likely to affiliate with similarly high-performing students. The powerful effects of peer influence could then be yet another factor contributing to the transformation of the intervention's short-term impact into long-term effects (Cohen and Prinstein, 2006; see also Hanuschek et al., 2006).

As a consequence of the impact of these processes, the social identity of minority students receiving the intervention may become even less of a source of concern. Psychological intervention in this sense is not at all small, as its effects can often be reinforced by the powerful self-validating nature of perception, motivation, and performance.

How can psychological interventions be transformed into practices that can be implemented throughout a school, a district, or a nation? Scaling up interventions into pedagogical practices suitable for widespread dissemination constitutes a substantial scientific endeavor. Several empirical questions immediately present themselves. For instance, will intervention effects be generalizable, or will they be primarily moderated by important features of the context, such as its racial composition (Cohen and Steele, 2002)? Social identity threat appears to be more acute when people constitute a numerical minority (Inzlicht and Ben-Zeev, 2000). An implication following from this, although speculative, is that interventions aimed at lessening such threat may be relatively more effective in institutions with a significant number of White and other nonstereotyped individuals. Will teachers be able to administer the interventions independently without the input of researchers with equal success? Experimental trials often try to minimize practitioners' and beneficiaries' awareness of the purpose of an intervention to protect the experiment's validity. But when an intervention is scaled up, its purpose and underlying rationale often become widely known. How is the effectiveness of a psychological intervention affected by students' or teachers' being aware of its purpose?

In an effort to address such questions in order to reach our aim of turning social-psychological interventions into widespread educational practices, we have continued working at our original school site with sustained success (Cohen et al., 2009). Moreover, we have expanded the project to include another school site with a more economically disadvantaged and predominately Latino American student body, where we have also obtained positive results. This is encouraging given that Latino Americans constitute the fastest growing minority group in the United States.

General Lessons about Intervention: Changing a Tension System

We now turn to some general observations that emerge from the consideration of social tension systems, which, among other things, are interactive in nature and constituted by social and psychological factors that can often be difficult to identify. If we are to maximize the possibility of bringing positive change to the classroom and other settings, it is imperative that we increase our understanding of how the factors making up a particular tension system can be enlisted in the process of creating and implementing interventions across a range of domains. The outcomes of interest could not only include academic outcomes, the focus of our discussion, but, among others, those related to health, well-being, and conflict (Boehm and Lyubomirsky, 2009; Cohen et al., 2007).

Sometimes Small Things Matter

A theme that emerges in the research summarized both in this chapter and in other chapters in this volume is that seemingly small interventions can have large effects when they target important social-psychological processes (Benartzi, Peleg, and Thaler, this volume; Thaler and Sunstein, 2008; Wansink, this volume). This is not a new idea because much of what made classical research in social psychology so noteworthy is that it demonstrated how seemingly subtle factors could have long-term effects. When these factors alter people's underlying values, attitudes, or self-concepts, those effects are particularly likely to persist (Freedman, 1965). This is especially true when these factors set in motion recursive cycles that can carry forward, and even augment, short-term effects (Cohen et al., 2009).

The notion that subtle shifts in psychological functioning can have considerable effects on important social outcomes can be seen not only in education but also in other domains, such as that of health. For instance, Pennebaker and his colleagues have consistently shown that having individuals engage in expressive writing requiring them to reflect on their thoughts and feelings related to a stressor in their lives can reduce stress. This in turn can improve health outcomes, even among cancer survivors and HIV+ patients (Petrie et al., 2004). Self-affirmation seems

to underlie some of these health benefits (Creswell et al., 2007).

No Intervention Is an Island

A corollary of the notion that small things matter is the idea that the effects of an intervention can, in turn, depend on contextual factors that can be obvious or subtle (Bertrand et al., 2005). One critical implication of that concept is that the impact of any intervention will depend on the forces already at play in a given social environment. Interventions should not be thought of independently from the context in which they are administered (Bertrand et al., 2005). Although patently obvious, this point is often underappreciated or even ignored. Social policy, including that involving education, is replete with instances, for example, the educational policy geared towards the reduction of class size. In response to educational research showing a negative relationship between class size and academic performance, well-intentioned policy makers enacted initiatives designed to reduce the number of students in classes. However, implementing these initiatives could, at least initially, require employing less well trained and less experienced teachers, even though a lack of teacher training and experience is associated with negative academic outcomes for students. At least in the short term, the implementation of these initiatives could put at risk any potential gains that would result from a reduction in class size, and in turn, having any number of negative outcomes, including the waste of scarce resources and the rejection of a potentially useful strategy for improving student performance.

Any initiative undertaken to alter outcomes in a social environment must interact with preexisting elements in such a way that permits it to have its desired end. Moreover, as highlighted in our discussion of recursive cycles, the outcomes of such interventions may take time before they become apparent. Again, these notions carry several implications for policy makers. In our hypothetical situation, they could give rise to two pragmatic implications. The first would be that class-size effects are based on the assumption that all other factors in the classroom are kept more or less constant, so provision for such constancy should be made in the implementation of the initiative. The second is the possibility that an intervention's real impact may not be observed until a significant amount of time has passed, so sufficient resources should be provided to allow for a fair test of the intervention's effectiveness. For instance, in our example, this would involve waiting until a sufficient number of trained and experienced teachers are produced or recruited.

Another implication of the idea that intervention effects can depend on the contextual factors is that the impact of interventions can appear disproportionately large given the resources and time dedicated to them. This is what we believe occurred in the case of our affirmation intervention. Such an apparently disproportionate effect is contingent on existing factors that facilitate motivation and performance. Without adequately trained and committed teachers, sufficient material resources, social support, and students who have acquired the skills to perform better, psychological interventions stand little or no chance of having a significant impact of any size. For example, although our affirmation intervention might lead a student who does not know how to spell to have a more positive sense of self-integrity in the face of his or her inability to spell, it will not suddenly turn this student into an adequate speller. Moreover, psychological interventions might prove less effective in a disadvantaged school where students may have been consistently exposed to less qualified teachers and had fewer resources dedicated to them over time than in a middle-class school.

However, when such resources are present, psychological interventions can catalyze their impact (Cohen et al., 2006, 2009; Menec et al., 2006), and lead to a situation in which an intervention's effects seem unusually large or influential. What appears to be a small or brief event if viewed in isolation acts as a catalyst for a process that realigns the elements in the environment so as to allow positive conditions, which were not previously fully realized, to manifest their impact more completely. For example, critical feedback had a strong and positive impact on stereotyped students' performance, but only when accompanied with a message that ascribed the rigor of the feedback to the evaluator's high standards and belief in the student's potential (Cohen and Steele, 2002). When the identity threat was alleviated, the learning resources could assert their full impact.

Look Before You Intervene and Above All Do Not Oversimplify

The fact that key outcomes in tension systems can rarely, if ever, be attributed to a single factor carries with it still another implication, that is, to question explanations and initiatives that seek to oversimplify the processes underlying intervention effects. In a classic article concerning this issue, Woodhead (1988) observed, "One of the problems in communicating the messages of [intervention research] is that the experimental design itself encourages disproportionate attention to be directed toward the critical manipulated variable as *the* cause of observed differences between experiment and control groups, no matter how remote in time or nature the outcome measures are

from the intervention" (p. 452). Focusing on a single cause can be an impediment to reaching the desired outcome because it can obscure our understanding of a social environment and keep us from addressing other critical factors in it.

Woodhead (1988) provided a concrete example of the potential dangers of ignoring this caution in his discussion of the effects of preschool interventions on long-term high school retention rates. He showed how their effects were mediated by other factors in the social environment. Early preschool interventions did produce a small gain in intellectual performance and student engagement in school when students began first grade. However, it was the positive impact that these gains had on the impressions of children held by teachers and by the school staff that made it less likely that the children would later be retained in a grade or be assigned to special education classes. Obviating these outcomes, in turn, made it more likely that the students would continue their education. The long-term impacts of the preschool interventions on later high school graduation rates, and even on postgraduation employment, were a result of how their effects interacted with other factors in the environment. In this case, the other factors were the perceptions of the students held by key "gatekeeper" individuals in the environment and the practices of holding low-achieving students back in a grade or assigning them to special education classes. Social context is key to understanding children's performance over time and the processes likely to impede or amplify the effectiveness of interventions (Bronfenbrenner, 1979).

Given the role that unobserved or underappreciated aspects of tension systems can have in producing critical outcomes, identifying the factors at work in an environment and examining how these interact with one another is essential to the creation of successful interventions. These activities increase the likelihood of developing strategies that can systematically alter the nature of the interactions taking place in a tension system so as to produce desired outcomes. For example, strategies could alter these interactions by introducing some new element into the environment or by changing the intensity of an existing factor in it. Clearly, a total or even comprehensive inventory of the factors making up a particular social environment is rarely, if ever, possible. Fortunately, based on our research findings, such an inventory is not necessary in order to effect significant and long-lasting change (Cohen et al., 2006). Although there are a multitude of factors at work in the majority of social environments, often only a few of them exercise a major role in producing critical outcomes, and still fewer are subject to manipulation.

For instance, during a careful observation of classrooms, researchers discovered a factor that exacerbated interracial antagonism in the classroom—competition over scarce resources, in particular the students' struggle for their teacher's attention and praise (Aronson and Patnoe, 1997). Given that competition can increase intergroup conflict and prejudice, the researchers reasoned that restructuring the classroom to facilitate more cooperative relationships between students could provide the basis for an effective intervention. The resulting jigsaw classroom, as their intervention was termed, accomplished exactly this. The children in a classroom were first separated into groups. Each child was then given a piece of the lesson plan to learn and to convey to others in his or her group. In order to learn the whole lesson plan, children were obliged to acknowledge and depend on others in their group regardless of their race or ethnicity. In other words, the intervention made it in the students' self-interest to cooperate with one another irrespective of each other's race or ethnicity. The jigsaw classroom creates a structure in which the processes leading to desired outcomes are more collective than individualistic, and as a consequence, intergroup antagonism is lessened. Although seemingly small, this intervention promoted positive intergroup relations by triggering processes that reduced what was often thought to be intractable long-term intergroup antagonisms.

Sometimes It Is Psychological

The observable level of student performance or other school-related behavior could be an inaccurate display of students' actual abilities. Indeed, Vygotsky (1978), the renowned education psychologist, introduced the construct "the zone of proximal development" to indicate the difference between a child's current level of performance and the level that he or she would be capable of attaining under optimal situational conditions. The restraining forces in an environment may depress students' willingness or ability to demonstrate their true ability. Underperformance can thus be characterized as an "ecological problem" (Cole and Bruner, 1971). Obviously, restraining forces can include objective impediments. An overcrowded classroom could lessen the likelihood that any individual student could be called upon to demonstrate what they know. However, there are also psychological factors that can act as restraining forces in such environments. A classic study showed that while young street vendors in Brazil were able to solve complex arithmetic problems in out-of-school settings, for instance rapidly adding up the price of several coconuts, they

failed to solve the same basic problem when it was presented on a written test in school (Carraher and Schliemann, 2002).

The label of "underachiever" captures the essence of such situations, because it implies that an individual has a level of skill that he or she is unwilling or unable to demonstrate. Work on test-anxiety has shown that the stress related to taking tests can impede performance, so much so that simply reducing their stress by removing testing time limits improves their performance to equal that of nonanxious students (Sarason, Mandler, and Craighill, 1952; see also Morris and Liebert, 1969). Paradoxically, it is sometimes those individuals who care most about performing well who are most unable to display their actual skill level when needed. This outcome is often characterized as choking under pressure in the "big game" or on a high-stakes standardized test. A similar phenomenon was found in seminal research that showed that the performance of low-income minority children on IQ tests and evaluative interviews was inhibited by psychological threat and "wariness" (Labov, 1970; Zigler, Abelson, and Seitz, 1973; Zigler and Butterfield, 1968; see also Cole and Bruner, 1971). Fortunately, there are other psychological factors that can mitigate such forces. In fact, this pivotal research also revealed that small procedural interventions that raise students' comfort in the test-taking situation, such as a friendly test proctor, can significantly increase these children's IQ scores and verbal fluency, sometimes dramatically.

As we stated earlier, it is critical to keep in mind that although a classroom or testing situation may appear to be the same for all those in it, this may not be the case. Due to differences in students' social identity and personal background it may have a radically different meaning, evoking different psychological reactions and apparently "objective" outcomes. As a consequence, in the words of Cole and Bruner, "it is not sufficient to use a simple equivalence-of-test procedure to make inferences about the competence of the two groups being compared" (1971, p. 871).

One Size Does Not Fit All

Our approach suggests the value of a targeted approach to psychological intervention. Like medical treatments, psychology-based interventions should ideally only be given to those needing them and who will benefit from them. This should be done not only to make the most effective use of time and material, but more important, to minimize the possibility of unforeseen adverse consequences. More generally, some interventions may prove less effective than others, and scaling them up before conducting a small-scale pilot study could not only waste resources and time, but also yield unforeseen negative consequences. For instance, attributional retraining can be ineffective when at-risk students receive poor instruction or lack the resources needed to improve (Menec et al., 2006). That situation may make it critical in disadvantaged areas to pair such interventions with skill-development workshops that provide students with the school resources they need (see Blackwell, Trzesniewski, and Dweck, 2007). Furthermore, the message of optimism that often surrounds such interventions may contradict students' actual experiences in the classroom, and lead to increasing frustration, disappointment, or mistrust (Wilson, Damiani, and Shelton, 2002). Interventions suggesting that the concerns of minority students are common and shared by majority-group members may be ineffective, and even counterproductive, when cues in institutional settings are continually reinstantiating identity threat in these students. For example, color-blind messages that downplay the importance of ethnicity can undermine minorities' trust and belonging when such messages are provided in the absence of actual institutional diversity, or when they convey that the positive distinctive qualities of one's culture will be ignored or should be suppressed (Purdie-Vaughns et al., 2008). In summary, psychological interventions will be more effective if the institutional setting provides adequate material and human resources. More generally, interventions need to be rigorously tested in any new context to monitor for unforeseen consequences, and ideally they should be given only to those who would benefit from them.

Timing Is Almost Always Important

The most critical aspect of an intervention can often be when it is administered, that is, its *timing*. Research on leadership offers an example with findings that show that a leader in work or school can change an organization's norms for the better, but only at certain junctures. Specifically, a leader's greatest impact occurs early in a project, prior to norms having been set; in the middle of the project, when groups naturally monitor their progress; and at the end, when group members take stock of the project (Hackman, 1998). A similar example is provided by research in early child education showing that interventions that target early childhood experiences, through preschool enrichment programs for instance, can have particularly high returns (Heckman, 2006).

The importance that *timing* can have in psychological interventions cannot be overstated. Psychological interventions, for instance, may be most effective

when administered at times of high stress as a means of interrupting a downward slide in functioning. In the educational domain, it could prove worthwhile to administer interventions at times of academic transition, such as those into middle school, high school, or college. These are times when the performance standards students are expected to meet shift upward, when their sense of identity is in flux, and their existing social support circles are disrupted. Each of these factors, alone or in concert, can heighten stress and feelings of exclusion. Intervening early in these transitions can have relatively larger benefits because they can interrupt recursive cycles triggered by such factors that would otherwise set students on a downward trajectory (Cohen et al., 2009).

It is also important to time an intervention to occur during the period in which it will have the most impact on an individual's *psychological environment*. If given too early, for instance, before students feel uneasy, the attributional retraining intervention could set off the very concerns it is intended to alleviate. It could, by suggesting to students that they *should* be wondering about their ability and belonging, make these thoughts salient when they otherwise would not have been (Wilson, Damiani, and Shelton, 2002; see also Pennebaker, 2001). Similarly, counting one's blessings or engaging in altruistic acts, activities that are often part of strategies designed to increase people's happiness, can be rendered ineffective by subtle changes in their timing or frequency (Boehm and Lyubomirsky, 2009).

One implication issuing out of the importance that timing can have in the development of interventions is the necessity of being able to identify not only *who* needs an intervention, but also *when* it is most needed. As in medical science, because the effects of psychological interventions can be harmful, unintended, or simply ineffective for certain individuals, it is as a general rule inadvisable to administer interventions indiscriminately. Likewise, for many of the same reasons, as well as others, it is inadvisable to administer an intervention too often, not often enough, or at times when it is inappropriate or irrelevant. Given this, in developing an intervention it is often critical to create methods for determining who needs it and when they need it. For instance, in our affirmation intervention research we have used, in conjunction with the intervention itself, validated climate assessments designed to assess students' perception of the school environment, as well as their psychological state, at more or less regular intervals to aid us in administering the intervention in a more targeted manner and at the most appropriate time. It is even possible to micro-time psychological interventions to occur at moments of maximal need for a given individual. For instance, through mobile technology it is possible to deliver interventions to people as they go about their normal lives and to tailor the timing and content of the intervention to each person's distinctive experiences and needs (see Heron and Smyth, 2010). We hope that ultimately practitioners and researchers will be able to apply psychology-based interventions in the way that physicians intervene medically. They will use a body of scientific research knowledge and its associated diagnostic technologies to help identify who should receive a treatment and when they should receive it.

Conclusion

The obvious but often overlooked notion that social environments such as schools and classrooms are complex tension systems composed of interacting factors, including recursive psychological processes, deserves the attention of researchers, practitioners, and policy makers. So does the idea that timely interventions, of whatever duration and magnitude, that address people's need for meaning, self-integrity, and belonging can have large and long-lasting effects on behavior and attitudes. Because of the interactive nature of social environments, an intervention's duration and magnitude depend on how it interacts with important processes existing in the environment. As a consequence, although a particular structural factor or set of factors—such as small class size, qualified teachers, or adequate funding—may be necessary to produce optimal outcomes, they may not be sufficient. Other factors in the environment, such as psychological processes, may suppress or obscure their impact.

In our view, unappreciated psychological factors have led to the questioning of the role of structural factors in schools, such as small class size and the degree of funding, in student achievement (see Heckman, Layne-Farrar, and Todd, 1996; for a review, see Burtless, 1996). However, as the research highlighted in this chapter shows, the introduction of new factors into a social environment, or the changing of preexisting ones, can make it more likely that such structural factors will exert their full impact (Lewin, 1951; Ross and Nisbett, 1991). For instance, the systematic introduction into the classroom of a psychological factor that was new, or if present already of relatively low intensity, increased students' ability or desire to avail themselves of the learning resources in the environment and their willingness or ability to demonstrate the skills and knowledge they had acquired (Blackwell, Trzesniewski, and Dweck, 2007; Cohen et al., 2006, 2009; see also Cohen and Steele, 2002; Cohen, Steele, and Ross, 1999; Menec et al.,

2006). The psychological factor introduced by the intervention catalyzed the impact of existing structural and material resources in a way that was subsequently expressed by students' improved motivation and performance.

The experiences, insights, and wisdom of the individuals intimately involved with a particular social environment must play a critical part in the scientific endeavor of determining if, when, and how interventions, including psychological interventions, can be made systematically effective. Already many educators, as well as others in the educational community, regularly use psychological strategies in their daily practice, often intuitively. For instance, among the many examples that exist (see Cose, 1997), some teachers have found that expressive writing, in which at-risk children associate their troubles with important values and literary stories, can have dramatic positive effects on students' engagement with school (Freedom Writers and Gruwell, 1999). The teacher portrayed in the movie *Stand and Deliver* and in the book by Mathews (1988), Jaime Escalante, found that challenging urban minority students with high academic standards and providing them with intensive support to reach those standards led them to earn achievement test scores as high as their more privileged White peers. Such examples convince us that partnerships of equals between practitioners and scientists hold the greatest promise for the development and implementation of psychological interventions of long-lasting and widespread impact.

Notes

We thank Sarah Wert and Eden Davis for comments on an earlier draft. Portions of the authors' research cited in this article were supported primarily by grants from the National Science Foundation's Research and Evaluation on Education in Science and Engineering program, the W. T. Grant Foundation, and the Spencer Foundation. Additional support was provided by the Russell Sage Foundation and the Nellie Mae Education Foundation.

References

Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin*, *97*, 129–133.

Aronson, E., and Patnoe, S. (1997). *The jigsaw classroom: Building cooperation in the classroom* (2nd ed.). New York: Addison Wesley Longman.

Aronson, J. (2002). Stereotype threat: Contending and coping with unnerving expectations. In J. Aronson (Ed.), *Improving academic achievement: Impact of psychological factors on education* (pp. 303–328). San Diego, CA: Academic Press.

Aronson, J., Fried, C. B., and Good, C. (2002). Reducing the effects of stereotype threat on African American college students by shaping theories of intelligence. *Journal of Experimental Social Psychology*, *38*, 113–125.

Aronson, J., and Inzlicht, M. (2004). The ups and downs of attributional ambiguity: Stereotype vulnerability and the academic self-knowledge of African American college students. *Psychological Science*, *12*, 829–836.

Aronson, J., Lustina, M., Keough, K., Brown, J. L., and Steele, C. M. (1999). When White men can't do math: Necessary and sufficient factors in stereotype threat. *Journal of Experimental Social Psychology*, *35*, 29–46.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice-Hall.

Ben-Zeev, T., Fein, S., and Inzlicht, M. (2004). Arousal and stereotype threat. *Journal of Experimental Social Psychology*, *41*, 174–181.

Bertrand, M., Karlan, D., Mullainathan, S., Shafir, E., and Zinman, J. (2005). *What's psychology worth? A field experiment in the consumer credit market.* NBER Working Paper No. 11892. National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w11892

Blackwell, L., Trzesniewski, K., and Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development*, *78*, 246–263.

Boehm, J. K., and Lyubomirsky, S. (2009). The promise of sustainable happiness. In S. J. Lopez (Ed.), *Handbook of positive psychology* (2nd ed., pp. 667–677). Oxford: Oxford University Press.

Bowen, W. G., and Bok, D. (1998). *The shape of the river: Long-term consequences of considering race in college and university admissions.* Princeton, NJ: Princeton University Press.

Branscombe, N. R., Schmitt, M. T., and Harvey, R. D. (1999). Perceiving pervasive discrimination among African Americans: Implications for group identification and well-being. *Journal of Personality and Social Psychology*, *77*, 135–149.

Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and design*. Cambridge, MA: Harvard University Press.

Brooks-Gunn, J. and Furstenburg, F. F. (1986). The children of adolescent mothers: Physical, academic, and psychological outcomes. *Developmental Review*, *6*(3), 224–251.

Burtless, G. (1996). *Does money matter? The effect of school resources on student achievement and adult success.* Washington, DC: Brookings Institution Press.

Carraher, D. W., and Schliemann, A. D. (2002). Is everyday mathematics truly relevant to mathematics education? In J. Moshkovich and M. Brenner (Eds.), *Monographs of the Journal for Research in Mathematics Education: Vol. 11. Everyday and academic mathematics in the classroom* (pp. 131–153). Reston, VA: National Council of Teachers of Mathematics.

Ceci, S. J., and Papierno, P. B. (2005). The rhetoric and reality of gap closing: When the "have-nots" gain but the "haves" gain even more. *American Psychologist*, *2*, 149–160.

Cohen, G. L., and Garcia, J. (2005). I am us: Negative stereotypes as collective threats. *Journal of Personality and Social Psychology*, *89*, 566–582.

———. (2008). Identity, belonging, and achievement: A model, interventions, implications. *Current Directions in Psychological Science*, *17*, 365–369.

Cohen, G. L., Garcia, J., Apfel, N., and Master, A. (2006). Reducing the racial achievement gap: A social-psychological intervention. *Science*, *313*, 1307–1310.

Cohen, G. L., Garcia, J., Purdie-Vaughns, V., Apfel, N., and Brzustoski, P. (2009). Recursive processes in self-affirmation: Intervening to close the minority achievement gap. *Science*, *324*, 400–403.

Cohen, G. L., and Prinstein, M. J. (2006). Peer contagion of aggression and health-risk behavior among adolescent males: An experimental investigation of effects on public conduct and private attitudes. *Child Development*, *77*, 967–983.

Cohen, G. L., Sherman, D. K., Bastardi, A., Hsu, L., McGoey, M., and Ross, L. (2007). Bridging the partisan divide: Self-affirmation reduces ideological closed-mindedness and inflexibility in negotiation. *Journal of Personality and Social Psychology*, *93*, 415–430.

Cohen, G. L., and Steele, C. M. (2002). A barrier of mistrust: How negative stereotypes affect cross-race mentoring. In J. Aronson (Ed.), *Improving academic achievement: Impact of psychological factors on education* (pp. 303–328). San Diego, CA: Academic Press.

Cohen, G. L., Steele, C. M., and Ross, L. D. (1999). The mentor's dilemma: Providing critical feedback across the racial divide. *Personality and Social Psychology Bulletin*, *25*, 1302–1318.

Cole, M., and Bruner, J. S. (1971). Cultural differences and inferences about psychological processes. *American Psychologist*, *26*, 867–876.

Cose, E. (1997). *Color-blind: Seeing beyond race in a race-obsessed world*. New York: HarperCollins.

Creswell, J. D., Lam, S., Stanton, A. L., Taylor, S. E., Bower, J. E., and Sherman, D. K. (2007). Does self-affirmation, cognitive processing, or discovery of meaning explain cancer-related health benefits of expressive writing? *Personality and Social Psychology Bulletin*, *33*, 238–250.

Creswell, J. D., Welch, W., Taylor, S. E., Sherman, D. K., Gruenewald, T., and Mann, T. (2005). Affirmation of personal values buffers neuroendocrine and psychological stress responses. *Psychological Science*, *16*, 846–851.

Crocker, J., and Major, B. (1989). Social stigma and self-esteem: The self-protective properties of stigma. *Psychological Review*, *96*, 608–630.

Croizet, J., and Claire, T. (1998). Extending the concept of stereotype threat to social class: The intellectual underperformance of students from low socioeconomic backgrounds. *Personality and Social Psychology Bulletin*, *24*, 588–594.

Darley, J. M. and Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, *44*, 20–33.

Davies, P. G., Spencer, S. J., and Steele, C. M. (2005). Clearing the air: Identity safety moderates the effect of stereotype threat on women's leadership aspirations. *Journal of Personality and Social Psychology*, *88*, 276–287.

Dholakia, U. M., and Morwitz, V. G. (2002). The scope and persistence of mere measurement effects: Evidence from a field study of consumer satisfaction. *Journal of Consumer Research*, *29*, 159–167.

Dillon, S. (2006, November 20). Schools slow in closing gaps between races. *New York Times*, p. A1.

Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality and development*. Philadelphia: Taylor and Francis/Psychology Press.

Dweck, C. S., Davidson, W., Nelson, S., and Enna, B. (1978). Sex differences in learned helplessness: (II) The contingencies of evaluative feedback in the classroom and (III) An experimental analysis. *Developmental Psychology*, *14*, 268–276.

Eccles, J. S., Lord, S., and Midgley, C. (1991). What are we doing to early adolescents? The impact of educational contexts on early adolescents. *American Journal of Education*, *8*, 520–542 .

Fiske, S., and Taylor, S. (1991). *Social cognition*. New York: McGraw-Hill.

Frable, D.E.S., Blackstone, T., and Scherbaum, C. (1990). Marginal and mindful: Deviants in social interactions. *Journal of Personality and Social Psychology*, *59*, 140–149.

Freedman, J. L. (1965). Long-term behavioral effects of cognitive dissonance. *Journal of Experimental Social Psychology*, *1*, 145–155.

Freedom Writers and Gruwell, E. (1999). *The Freedom Writers diary*. New York: Broadway Books.

Fryer, R. G., and Torelli, P. (2005). *An empirical analysis of "acting white."* NBER Working Paper No. 11334. National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w11334

Fullilove, R. E., and Treisman, P. U. (1990). Mathematics achievement among African American undergraduates at the University of California, Berkeley: An evaluation of the Mathematics Workshop Program. *Journal of Negro Education*, *59*, 463–478.

Good, C., Aronson, J., and Inzlicht, M. (2003). Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Journal of Applied Developmental Psychology*, *24*, 645–662.

Gordon, E. W., and Lemons, M. P. (1997). An interactionist perspective on the genesis of intelligence. In R. J. Sternberg and E. L. Grigorenko (Eds.), *Intelligence, heredity, and environment* (pp. 323–340). New York: Cambridge University Press.

Grimm, L. R., Markman, A. B., Maddox, T. W., and Baldwin, G. C. (2009). *Journal of Personality and Social Psychology*, *96*, 288–304.

Hacker, A. (1995). *Two nations: Black and White, separate, hostile, unequal*. New York: Ballantine.

Hackman, J. R. (1998). Why teams don't work. In R. S. Tindale, J. Edwards, and F. B. Bryant (Eds.), *Theory and research on small groups* (pp. 245–267). New York: Plenum.

Hanushek, E. A., Kain, J. F., Markman, J. M., and Rivkin, S. G. (2006). *Does peer ability affect student achievement?* NBER Working Paper No. 8502. National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w8502

Hart, B., and Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Paul H. Brookes Publishing Co.

Heckman, J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science*, *312*, 1900–1902.

Heckman, J., Layne-Farrar, A., and Todd, P. (1996). Does measured school quality really matter? An examination of the earnings-quality relationship. In G. Burtless (Ed.), *Does money matter? The effect of school resources on student achievement and adult success*. Washington, DC: Brookings Institution Press.

Henderson-King, E. I., and Nisbett, R. E. (1996). Antiblack prejudice as a function of exposure to the negative behavior of a single Black person. *Journal of Personality and Social Psychology*, *71*, 654–664.

Heron, K. E., and Smyth, J. M. (2010). Ecological momentary interventions: Incorporating mobile technology in psychosocial and health behavior treatments. *British Journal of Health Psychology*, *15*, 1–39.

Huo, Y. J., Smith, H. J., Tyler, T. R., and Lind, E. A. (1996). Superordinate identification, subgroup identification, and justice concerns: Is separatism the problem, is assimilation the answer? *Psychological Science*, *7*, 40–45.

Inzlicht, M. and Ben-Zeev, T. (2000) A threatening intellectual environment: Why females are susceptible to experiencing problem-solving deficits in the presence of males. *Psychological Science*, *11*, 365–371.

Jencks, C., and Phillips, M. (1998). *The Black-White test score gap*. Washington, D. C.: The Brookings Institution.

Jussim, L., and Harber, K. (2005) Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review*, *9*,131–155.

Kaiser, C. R., Brooke, V., and Major, B. (2006). Prejudice expectations moderate preconscious attention to cues that are threatening to social identity. *Psychological Science*, *17*, 332–338.

Kleck, R. E., and Strenta, A. (1980). Perceptions of the impact of negatively valued characteristics on social interaction. *Journal of Personality and Social Psychology*, *39*, 861–873.

Lavob, W. (1970). The study of language in its social context, *Stadium Generale*, *23*, 30–87.

Lazarus, R. S., and Cohen, J. B. (1977). Environmental stress. In I. Altman and J. F. Wohlwill (Eds.), *Human behavior and environment*. (Vol. 2, pp. 222–230) New York: Plenum.

Leary, M. R., and Baumeister, R. F. (2000). The nature and function of self-esteem: Sociometer theory. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 32, pp. 1–62). San Diego, CA: Academic Press.

Lewin, K. (1948) *Resolving social conflicts: Selected papers on group dynamics*. G. W. Lewin (Ed.). New York: Harper and Row.

———. (1951) *Field theory in social science: Selected theoretical papers*. D. Cartwright (Ed.). New York: Harper and Row.

Lockwood, P., and Kunda, Z. (1997). Superstars and me: Predicting the impact of role models on the self. *Journal of Personality and Social Psychology*, *73*, 91–103.

Martens, A., Johns, M., Greenberg, J., and Schimel, J. (2006). Combating stereotype threat: The effect of self-affirmation on women's intellectual performance. *Journal of Experimental Psychology*, *42*, 236–243.

Marx, D. M., Brown, J. L., and Steele, C. M. (1999). Allport's legacy and the situational press of stereotypes, *Journal of Social Issues*, *55*, 491–502.

Mathews, J. (1988). *Escalante: The best teacher in America*. New York: Henry Holt and Company.

Menec, V. H., Perry, R. P., Struthers, C. W., Schonwetter, D. J., Hechter, F. J., and Eichholz, B. L. (2006). Assisting at-risk college students with attributional retraining and effective teaching. *Journal of Applied Social Psychology*, *24*, 675–701.

Morris, L. W., and Liebert, R. M. (1969). Effects of anxiety on timed and untimed intelligence tests. *Journal of Consulting and Clinical Psychology*, *33*, 240–244.

Muennig, P., and Woolf, S. H. (2007). Health and economic benefits of reducing the number of students per classroom in US primary schools. *American Journal of Public Health*, 97, 2020–2027.

Neal, D. A. (2005). *Why has black-white skill convergence stopped?* NBER Working Paper No. W11090. National

Bureau of Economic Research. Retrieved from http://ssrn.com/abstract=657602

Nisbett, R. E. (2009). *Intelligence and how to get it: Why schools and cultures count.* New York: W. W. Norton and Co.

O'Brien, L. T., and Crandall, C. S. (2003). Stereotype threat and arousal: Effects on women's math performance. *Personality and Social Psychology Bulletin*, *29*, 782–789.

Pennebaker, J. (2001). Dealing with a traumatic experience immediately after it occurs. *Advances in Mind-Body Medicine*, *17*, 160–162.

Petrie, K. J., Fontanilla, I., Thomas, M. G., Booth, R. J., and Pennebaker, J. W. (2004). Effect of written emotional expression on immune function in patients with human immunodeficiency virus infection: A randomized trial. *Psychosomatic Medicine*, *66*, 272–275.

Purdie-Vaughns, V., Steele, C., Davies, P., Ditlmann, R., and Randall-Crosby, J. (2008). Social identity contingencies: How diversity cues signal threat or safety for African Americans in mainstream institutions. *Journal of Personality and Social Psychology*, *94*, 615–630.

Richards, Z., and Hewstone, M. (2001). Subtyping and subgrouping: Processes for the prevention and promotion of stereotype change. *Personality and Social Psychology Review*, *5*, 52–73.

Rosenthal, R., and Jacobson, L. (1992). *Pygmalion in the classroom* (Expanded edition). New York: Irvington.

Ross, L., and Nisbett, R. E. (1991). *The person and the situation.* Philadelphia: Temple University Press.

Rydell, R. J., McConnell, A. R., and Beilock, S. L. (2009). Multiple social identities and stereotype threat: Imbalance, accessibility, and working memory. *Journal of Personality and Social Psychology*, *96*, 949–966.

Sanbonmatsu, L., Kling, J. R., Duncan, G. J., Brooks-Gunn, J. (2006). *Neighborhoods and academic achievement: Results from the Moving to Opportunity experiment.* NBER Working Paper No. 11909. National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w11909

Sarason, S. B., Mandler, G., and Craighill, P. G. (1952). The effect of differential instructions on anxiety and learning. *Journal of Experimental Psychology*, *59*, 185–191.

Schmader, T., and Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology*, *85*, 440–452.

Schmader, T. , Johns, M., and Forbes, C. (2008). An integrated process model of stereotype threat effects on performance. *Psychological Review*, *115*, 336–356.

Schmader, T., Major, B., and Gramzow, R. H. (2001). Coping with ethnic stereotypes in the academic domain: Perceived injustice and psychological disengagement. *Journal of Social Issues*, *57*, 93–111.

Shapiro, J. R., and Neuberg, S. L. (2007). From stereotype threat to stereotype threats: Implications of a multi-threat framework for moderators, mediators, and interventions. *Personality and Social Psychology Review*, *11*, 107–130.

Sherman, D. K., and Cohen, G. L. (2006). The psychology of self-defense: Self-affirmation theory. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 38, pp. 183–242). San Diego, CA: Academic Press.

Simmons, R. G., Black, A., and Zhou, Y. (1991). African-Americans versus White children and the transition into junior high school. *American Journal of Education*, *99*, 481–520.

Spencer, S. J., Steele, C. M., and Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*, 4–28.

Steele, C. M. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 21, pp. 261–302). New York: Academic Press.

———. (1997). A threat in the air: How stereotypes shape the intellectual identities and performance of women and African-Americans. *American Psychologist, 52*, 613–629.

Steele, C. M., and Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*, 797–811.

Steele, C. M., Spencer, S. J., and Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 34, pp. 379–440). San Diego, CA: Academic Press.

Steele, C. M., Spencer, S., Nisbett, R., Hummel, M., Harber, K., and Schoem, D. (2004). *African American college achievement: A "wise" intervention.* Manuscript submitted for publication.

Stone, J., Lynch, C. I., Sjomeling, M., and Darley, J. M. (1999). Stereotype threat effects on Black and White athletic performance. *Journal of Personality and Social Psychology*, *77*, 1213–1227.

Thaler, R. H., and Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness.* New York: Penguin Books.

Tierney, R. J., Crumpler, T. P., Bertelsen, C. D., and Bond, E. L. (2003). *Interactive assessment: Teachers, parents, and students as partners.* Norwood, MA: Christopher-Gordon.

Tyler , T. R. (2004). Procedural justice. In A. Sarat (Ed.), *The Blackwell companion to law and society* (pp. 435–452). Malden , MA : Blackwell.

U. S. Department of Education National Center for Education Statistics (2009). *The condition of education* (NCES 2009-081). Washington, D.C.

Vygotsky, L. S. (1978). *Mind and society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Walton, G. M., and Cohen, G. L. (2007). A question of belonging: Race, social fit, and achievement. *Journal of Personality and Social Psychology, 92*, 82–96.

Walton, G. M. and Spencer, S. J. (2009). Latent ability: Grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students. *Psychological Science*, *20*, 1132–1139.

Wilson, T. D. (2006). The power of social-psychological interventions. *Science*, *313*, 1251–1252.

Wilson, T. D., Damiani, M., and Shelton, N. (2002). Improving the academic performance of college students with brief attributional interventions. In J. Aronson (Ed.), *Improving academic achievement: Impact of psychological factors on education*. San Diego, CA: Academic Press.

Wilson, T. D., and Linville, P. W. (1985). Improving the performance of college freshmen with attributional techniques. *Journal of Personality and Social Psychology*, *49*, 287–293.

Woodhead, M. (1988). When psychology informs public policy: The case of early childhood intervention. *American Psychologist*, *43*, 443–454.

Zigler, E., Abelson, W. D., and Seitz, V. (1973). Motivational factors in the performance of economically disadvantaged children on the Peabody Picture Vocabulary Test. *Child Development*, *44*, 294–303.

Zigler, E., and Butterfield, E. C. (1968). Motivational aspects of changes in IQ test performance of culturally deprived nursery school children. *Child Development*, *39*, 1–14.

CHAPTER 20

Beyond Comprehension

Figuring Out Whether Decision Aids Improve People's Decisions

PETER UBEL

A married couple in their mid-40s with two young children ask their financial advisor whether they should increase the percentage of their assets placed into high-yield stocks. A seventeen-year-old high school student meets with her guidance counselor for advice on where to apply to college. A sixty-five-year-old man with a nonmetastatic prostate cancer asks his physician whether he should have his prostate surgically removed.

Each of these people is facing what is known as a *preference sensitive decision*, where the right choice depends in part on that person's specific preferences (O'Connor et al., 1999). The best investment choice, for instance, depends on a given person's risk tolerance; the best college choice depends on a person's preferences for big cities versus small towns, liberal arts versus engineering classes; and the best approach to prostate cancer will depend on how concerned a man is about the risk of impotence or incontinence from treatment.

In each of these cases, the person making the decision is looking for help from a neutral party—someone who can help them make a decision that is consistent with their underlying goals and preferences. Which raises an important question, How do decision counselors know when they have improved people's decisions?

This question is important because, when left to their own devices, people will not always make the right decision. That is one of the reasons why people seek advice from lawyers, counselors, physicians, and financial advisors—they recognize that they will have a hard time becoming adequately informed about the issues relevant to the decision at hand and therefore turn to experts who can help them. In the face of important, preference-sensitive decisions, the job of a decision counselor should be to help decision makers comprehend information about their alternatives and integrate their information with their individual preferences.

In this article, however, I will show why decision counselors need to go beyond helping people comprehend their decision alternatives. The field of judgment and decision making is replete with examples of people who comprehend their decision alternatives and nevertheless make bad decisions.

I will discuss these important issues in the context of medical decision making, a setting in which decisions often have unusually high stakes, involve complex choice sets, and typically do not provide the decision makers, the patients, with the ability to fully inform themselves about their alternatives. The decision makers, in other words, need help making the decisions. I will specifically focus on a growing movement within medicine to provide patients with decision aids (DAs)—structural educational materials designed to inform people about their decision alternatives. While the context of this paper will focus on medical decision making, the lessons I draw from medical decision making, and the preliminary criteria I develop for judging when a decision has been improved, are relevant in other domains where people face high-stakes decisions and need help sorting through complex information.

My goal in this article is to explain why the criteria that experts on shared decision making have been using to test DAs in health contexts—that is, whether DAs increase knowledge and reduce decisional conflict—are inadequate for determining whether a given DA actually improves people's health-care decisions. In critiquing these criteria, I will evaluate the strengths and weaknesses of seven additional criteria. I conclude that no single criterion is sufficient for evaluating a DA but, instead, that we need to utilize a broad array of testing standards in order to judge whether a specific DA improves people's decisions.

Healthcare Decision Aids: Structure and Evaluation

Recognizing that physicians are not always effective decision coaches, a movement has grown within health care to supplement physician communication with DAs (Bekker et al., 1999; Molenaar et al., 2000; O'Connor et al., 2009). These DAs are patient-education materials that have been informed by decision analysis and that structure information in ways that make patients aware of the trade-offs inherent in their treatment choices—for instance, explaining the possible outcomes of treatment A versus treatment B and the likelihood of each outcome. These DAs also strive to activate patients, showing them the important role that their own preferences should play in determining their treatment choice.

Decision aids have been rigorously tested in randomized trials and have been shown to increase patient knowledge and satisfaction with their decision while reducing decisional conflict (O'Conner et al., 1999). Indeed, DAs are typically judged as effective or ineffective in large part based on these criteria. These criteria, then, are a good starting point for any assessment of how to help people make good decisions.

Knowledge and Comprehension

Health-care DA developers have placed great emphasis on testing whether DAs adequately inform patients about their health-care alternatives. Indeed, DAs are evaluated during their development for balance, thoroughness, and comprehensibility. Constructing an informative and comprehensible DA is often challenging, forcing the DA developers to make difficult judgments about how much information to include, how to help people understand probabilistic outcome data, and how to engage people in the information without overwhelming them or boring them. Often DA developers refine the materials through focus groups and cognitive interviews. The best DAs are even pretested by literacy experts to make sure they are not written above a seventh-grade reading level. The end result, typically, is a high-quality product that significantly increases patients' knowledge of their health-care circumstances and their treatment alternatives.

Decisional Conflict and Decision Satisfaction

Health-care DA developers have also contended that DAs should increase decision satisfaction while reducing decisional conflict (O'Connor et al., 1999). They define decisional conflict as "the uncertainty about which course of action to take when choice among competing actions involves risk, loss, regret, or challenge to personal life values" (O'Conner, 1995). Signs of decisional conflict include "verbalized uncertainty, expressing concern about undesired outcomes, wavering between choices, delaying decisions, questioning personal values, being preoccupied with the decisions, and feeling emotionally distressed by the decision" (O'Connor, Jacobsen, and Stacey, 2002, p. 571). And they have developed a measure of such conflict, which they contend that a good DA will reduce (O'Connor, 1995). Along similar lines, developers of health-care DAs have promoted the idea that good DAs will increase patient satisfaction with the decisions they make (O'Connor et al., 1999) and again have developed a scale to measure such satisfaction (Holmes-Rovner et al., 1996).

Inadequacy of These Criteria

The shared-decision-making community in medicine has largely assumed that if you give decision makers freedom and information, they will experience high satisfaction with their decisions, will be unconflicted about their choices, and will make decisions that reflect stable underlying preferences, or values. A wealth of studies, however, have demonstrated that these assumptions are often false and that free and informed decisions are not always good decisions.

Knowledge Does Not Protect People from Bias

The judgment and decision-making literature is replete with evidence of biases that can unduly influence even the most informed decisions makers. I illustrate this problem with a study my research team conducted on a DA we designed to help women contemplate whether to take tamoxifen to prevent breast cancer. Tamoxifen is a hormonelike medication that was initially used in breast cancer patients to reduce the chance that the breast cancer will return. Clinical trials have more recently demonstrated that tamoxifen can be used in high-risk women, to prevent them from developing a *first* breast cancer (Day, 2001). For instance, a woman with a 6% chance of developing a first breast cancer in the next 5 years (based on things like family history and age) can cut that risk in half by taking tamoxifen. But this medication is not harmless. Women taking tamoxifen have a chance of developing blood clots, endometrial cancer, hot flashes, and cataracts. In short, the decision to take tamoxifen to prevent a first breast cancer is by everyone's reckoning a preference-sensitive decision.

In developing our DA, we wanted to test how much women's attitudes toward tamoxifen would

be influenced by subtle changes in how we presented them with information about its risks and benefits. For example, we varied the denominator we used to illustrate the frequency of tamoxifen's side effects. We informed some women that 17 out of 100 women taking tamoxifen would experience cataracts, and others that 170 out of 1000 women would experience this side effect. We also varied whether women learned first about the least common side effects of tamoxifen or the most common side effects. We found that women's attitudes toward tamoxifen significantly varied depending on these two subtle manipulations. Women were more concerned about tamoxifen's side effects when they were told how many out of a thousand women would experience these side effects and also more worried when the last side effect they learned about occurred with a high probability. Nevertheless, women's knowledge of tamoxifen's side effects was not influenced by either the denominator we chose in describing the risks nor the order in which we presented the risks (Zikmund-Fisher et al., 2008).

In a famous study, McNeil and colleagues discovered that people are more willing to undergo a surgical operation with a 90% survival rate than one with a 10% mortality rate (McNeil et al., 1982). Framing outcomes in terms of survival increased the desirability of the intervention without altering people's comprehension. Either way of framing the information—90% survival or 10% mortality—will lead to similar comprehension, while causing people to make different decisions, depending on whether the framing triggers people's aversion to losses (Tversky and Kahneman, 1981).

These studies illustrate an important challenge facing DA makers: seemingly neutral manners of presenting information can bias people's judgments and decisions, even if at the same time they increase people's knowledge of their decision alternatives. Knowledge, therefore, is a necessary part of any good decision, but it is not sufficient (Kennedy, 2003). A good DA should not only help people comprehend their choice alternatives but should also do so in a way that will minimize decisional biases.

Is Conflict Such a Bad Thing?

As mentioned above, health-care-decision experts also evaluate DAs according to whether they reduce decisional conflict. But their evaluation criterion rests on the assumption that reducing conflict ought to be a goal of DA. This assumption is questionable.

Consider two women deciding on early-stage cancer treatment. The first woman searches the Internet and finds a company website that explains why its pill is the best available treatment for this cancer. She asks her doctor to prescribe the pill. She feels comfortable about her decision and has no decisional conflict and high decision satisfaction. By contrast, the second woman receives access to a DA designed by a nonprofit foundation committed to helping people make informed choices. The DA provides information on several treatment options. With the help of this DA she learns about the risks and benefits of the treatments and picks the choice that she thinks fits her preferences best. But even though she has dismissed alternatives that she realizes would not suit her preferences well, she is still not sure she has made the right decision and feels conflicted about what the best one is.

This second woman has higher decisional conflict and lower satisfaction than the first and, therefore, by the standards that dominate the medical decision-making world, would be viewed as having used a *worse* DA. By contrast, the person who relied upon the industry information, which was quite persuasive, felt unconflicted about her decision. Thus, by this criterion, the industry "DA" would be judged superior to the less biased one. In judging DAs, we need to be open to the idea that a good decision may still leave people with substantial decisional conflict (Nelson et al., 2007).

Additional Criteria for Judging "Good" Decisions

Decision-making experts in health care have adopted criteria for judging DAs that follow closely upon models of rational choice. The shared-decision-making community in medicine has largely acted as if a free and informed decision is a good one, and therefore a DA that enables people to comprehend their health-care alternatives, and thereby reduce decisional conflict, is a good DA. But what other criteria might we consider for judging whether a DA has improved people's decisions? Below I will discuss seven additional criteria (table 20.1) and their role in evaluating whether an intervention like a DA has improved people's decisions.

The Expected Utility Criterion

If the goal of a DA is to help people make decisions consistent with their preferences, then a DA could theoretically accomplish this goal by quantifying people's preferences and calculating the expected utility of each decision alternative. If a decision analyst knew the utility that a prostate-cancer patient placed on impotence and incontinence, for instance, she could insert those utility values into a decision model and

Table 20.1 Criteria available for evaluating whether decision aid has improved decisions

Standard criteria used in health care
1. Increase in knowledge
2. Satisfaction with decision
3. Reduction in decisional conflict
New criteria discussed in this chapter
1. Maximization of expected utility
2. Reduction of mispredictions: for example, accurate beliefs about consequences
3. Increased happiness: that is, moment to moment moods
4. Invariance: decision less susceptible to nonnormative influences
5. Correlational validity: decision shifts appropriately with change in risks/benefits
6. Time: decision maker has adequate time to process alternatives
7. Adherence: decision maker follows through on decision

tell that man which treatment maximizes his expected utility.

This approach has the advantage of "doing the math" for people when they are faced with decisions that are too complex to otherwise grasp. Indeed, decision analysis has been proposed as a method for determining whether a given decision is preference sensitive (Ubel and Loewenstein, 1997). If altering utilities across plausible ranges does not alter which alternative has the highest expected utility, then the decision is not preference sensitive.

However, in cases where the utility values do matter, I am concerned that our ability to quantify utility values is often too imprecise to determine which alternative is best. For example, the most common method of measuring health-related utilities is the standard gamble method, which was derived from the axioms of von Neumann and Morgenstern (von Neumann and Morgenstern, 1947). In a standard gamble utility elicitation, a man might be asked what chance of death he would take to rid himself of impotence (Gold et al., 1996). A closely related alternative known as the time-tradeoff method would ask that man how many of his remaining years of life he would give up in order to avoid impotence (Torrance, 1976; Torrance, Thomas, and Sackett, 1972). These preference measures are extremely difficult for many people to understand. For instance, people who have difficulty with probabilities and frequencies are often confused by the questions and give nonsensical answers (Woloshin et al., 2001). People often also raise moral objections to these questions (Baron and Spranca, 1997). In time trade-off elicitations, it is common for people to say they would not give up any time to improve their health, even when considering horrendous health-care problems. They respond this way because they feel it would be wrong to give up any amount of their lives, the moral equivalent of committing suicide.

Utility elicitation measures are also potentially flawed because they are influenced by affective forecasting errors. Most members of the general public, for example, assume that physical disabilities would have a much larger impact on their emotional well-being than people with those disabilities report (Ubel et al., 2005). Therefore, even people who understand utility elicitations and who have no moral objections to the questions may nevertheless bring inaccurate beliefs about the health-state question to mind, thereby biasing their responses.

These affective forecasting errors would likely plague an alternative preference measurement method—conjoint analysis. In conjoint analysis, decision makers are given a series of pair-wise choices, with random variation of specific attributes for each choice (Green and Srinivasan, 1978). Across conjoint choices, it is possible to analyze the weight that any given attribute contributes to a person's decisions. These weights can then be plugged into a decision analysis as utility values. Conjoint analysis avoids some of the problems of standard gamble and time trade-off utility measures. The conjoint approach does not usually raise moral objections from participants and does not involve the use of difficult concepts like percentages. But the conjoint approach still opens the door to affective forecasting errors. If a person incorrectly predicts that a colostomy would make him miserable, for example, then the presence of a colostomy in one of the choice pairs will dominate the person's conjoint decisions. Conjoint analysis is also flawed because it

still requires people to make complex decisions. A conjoint analysis might ask people to compare two choices over five attributes. This type of decision is susceptible to a range of well-known decisional biases.

Thus, decision analytic models can help determine whether a decision is preference sensitive and can even help reveal which preferences are most influential in determining the best choice. But the measurement of preferences—that is, of health-related utilities—is often too imprecise for this criterion to point toward the "correct" choice.

Reduction of Mispredictions

As the previous discussion illustrated, mispredictions stand as formidable barriers to optimal decision making. People can mispredict how decision-relevant outcomes will affect their emotional lives (Ubel, 2006). For instance, when making investment decisions, people may seek out high risks in hopes of transforming their lives with a huge payout. Similarly, when deciding how to treat their inflammatory bowel disease, patients may inappropriately eliminate surgical options out of a mistaken belief that a colostomy would make them miserable (Smith et al., 2006). People can also mispredict the nonemotional consequences of specific circumstances. For example, people with kidney failure overestimate how much a successful kidney transplant will improve their job prospects (Smith et al., 2008). A well-designed DA should reduce or eliminate such forecasting errors. But what would it mean for a DA to do this?

At a minimum, a DA should provide people information about the emotional and nonemotional consequences of specific decision-related outcomes. This might sound trivial, but this approach has not historically been the norm for DA developers. Financial advisors might advise people on the chance of losing half their savings versus the chance of quadrupling their investment. But I expect that few such advisors provide clients with information on how happy they are likely to be with each of these outcomes. I expect most financial advisers do not even recognize the true relationship between net worth and happiness (Diener and Seligman, 2004). Similarly, in health-care DAs, patients are presented with neutral language describing specific health-related outcomes. But they are not typically given complete information about the consequences of these outcomes.

DAs could reduce mispredictions by giving people "the answers"—telling them, for instance, how happy people are who have experienced the circumstance in question (Gilbert et al., 2009). A financial advisor could give a client information about the average happiness of people with a net worth of $500,000 versus $2.5 million. A health-care DA could report on the average happiness level of people with and without kidney failure.

Rather than give people the answers, DAs could reduce mispredictions by using debiasing techniques that help people correct affective forecasting errors themselves. For example, Wilson and Gilbert asked people to write out a diary of what their days would be like following a college football loss, and in doing so discovered that people became more accurate at estimating their moods on those days (Wilson, Meyers, and Gilbert, 2003). Similarly, my colleagues and I asked people to think about how they have responded in the past to emotionally salient circumstances, and doing so reduced affective forecasting errors for experiencing a severe disability (Ubel, Loewenstein, and Jepson, 2005).

Thus, there are two ways to use DAs to reduce mispredictions. The challenge will be to determine whether either of these techniques works, a challenge that will not easily be met. Suppose, for instance, we tell decision makers that people with kidney failure are, on average, almost as happy as people with normal kidneys (Riis et al., 2005). They could respond to this information in one of several ways:

1. *Complete disbelief*: They might deny that people are really that happy, perhaps questioning whether people with kidney failure are giving honest answers when asked how happy they are or believing that the DA developers have an agenda to promote.
2. *Believe others but not self*: They might accept that most people with kidney failure are happy but not believe that *they* would be so happy.
3. *Total acceptance*: They might believe that they, like the average person, would largely adapt to having kidney failure.

If they respond in the first manner, with complete disbelief, then we can confidently conclude that we have *not* adequately debiased them from affective forecasting errors. In this case, we ought to do more to convince them or find yet other ways to debias them.

But what about the second type of response above, where a decision maker holds an accurate belief about how an average person responds to a given circumstance, but is convinced that he will respond differently? For any given individual, it is impossible to know whether the decision maker is right or wrong. Some people really *are* made miserable by circumstances to which the average person can adapt. The same goes for the third response above. A person who believes that they will adapt to a given circumstance because most people adapt to the circumstance may *still* be making an affective forecasting error. For all

we know, they won't adapt as much as the average person. In either of these cases, it is impossible to tell whether a specific individual has correctly predicted their response to a given circumstance.

Therefore, to judge debiasing efforts, DA developers need to assess *aggregate responses.* If most decision makers believe they will be the exception to the rule, then the DA developer has not debiased them. Ideally, the forecasts of decision makers will map, as a group, onto the actual reports of people experiencing the circumstances in question. Better yet, longitudinal studies could verify which ways of informing people about circumstances are best at mapping onto how they will actually respond.

In short, the evaluation of DAs should expand to test, in longitudinal studies, whether people who receive the DA are able to predict the emotional and nonemotional consequences of decision-relevant outcomes.

The Happiness Criterion

The flaws identified with these first two alternative criteria raise the possibility that our problem—of identifying when a third party has improved someone's decision—can be solved by resorting to a happiness criterion. Specifically, we could test whether people who receive a DA are happier than those who do not. This criterion is based on the grounds that people do not know what makes them happy or unhappy, so a third party should figure out what does make them happy and find a way to convince them to make decisions that maximize their happiness (Ubel, 2009). This criterion differs from the second criterion, the forecasting criterion, by judging decisions after the fact, rather than before. We know that a decision is better when it is based on the consequences of the decision rather than the process that led to the decision.

The happiness criterion has many important strengths. All else equal, people generally want to be happy rather than unhappy, preferring positive moods to negative ones. Yet people do not always manage, even when well informed and uncoerced, to make decisions that maximize these aspects of well-being (Ubel, 2006). Thus, it would seem to be a good thing for DAs to protect people from making decisions that reduce their happiness.

But the happiness criterion suffers from two major weaknesses. First, experts do not agree on how to define happiness. Some define happiness narrowly, as the balance of positive and negative affect (Bentham, 1907; Kahneman, Wakker, and Sarin, 1997). By this definition, happiness is quantifiable and can be used to judge the impact of specific circumstances, or even decisions, on people's emotional well-being.

But this hedonic view of happiness strikes many people as being too narrow (Griffin, 1989; Loewenstein and Ubel, 2008). People care about many aspects of their lives beyond their moment-to-moment mood. For instance, they care about freedom for freedom's sake, preferring to trade-off some amount of happiness to increase their freedom. In addition, people care about opportunities and capabilities independent of how any limits on opportunities or capabilities influence their mood (Sen, 2004). Thus, for example, even when people recognize that their happiness will not be significantly reduced by a loss of income or by a new disability, most will nonetheless desire to maintain their income and their physical functioning (Damschroder, Zikmund-Fisher, and Ubel, 2005). Sen, in fact, contends that capabilities matter in large part because people are so good at adapting, emotionally, to unjust circumstances (Sen, 2004). Slavery would not be tolerable even if slaves were happy. Kidney failure would not be inconsequential just because people with kidney failure managed to adapt.

In short, DAs should be evaluated to see if they increase people's overall sense of well-being, but we should also be aware of how DAs influence freedom and capabilities. We cannot assume that if a DA improves people's moods, it has therefore improved their decisions. Nor can we assume that if a DA reduces people's happiness, it has therefore influenced them to make bad decisions. Sometimes decision makers make decisions solely to promote *other people's* interests, occasionally sacrificing their happiness for the sake of others. We would not want to call these decisions misguided.

The Invariance Criterion

DAs could also be evaluated by the standard of invariance—the idea that decisions should not change when the pros and cons of the decision alternatives remain the same. By this criterion, if I favor surgery A when I learn it has a 90% survival rate, then I should also favor it when I discover it has a 10% mortality rate—because a 90% survival rate is equivalent to a 10% mortality rate. Similarly, if I decide to take a medication with a 3-in-100 chance of migraines, I should not change my mind when I discover that the risk is 30 in 1000.

My colleagues and I have had success developing several methods for eliminating the influence of these decisional inconsistencies in health-care DAs. For example, we discovered that graphical representations of probability information can reduce the inappropriate influence of anecdotes, what I refer to in my medical practice as "the Aunt Millie problem." Like

most clinicians, I have encountered patients who reject plausible treatment alternatives out of hand because of something they heard from a friend or relative. Such encounters suggest that the way people feel about risks and benefits can be influenced by anecdotal information.

We explored this phenomenon in a survey of prospective jurors in Philadelphia (Ubel, Jepson, and Baron, 2001). We asked people to imagine that they had chest pains from coronary artery disease and that there were two treatment alternatives to choose from: bypass surgery, which had a 75% chance of curing their chest pains but which required open-heart surgery and a prolonged recovery period; or balloon angioplasty, which had only a 50% chance of curing their chest pains but was a much less arduous procedure. We illustrated this choice with a series of uninformative anecdotes, relaying the stories of hypothetical people who had received each treatment and had either experienced a cure of their chest pain or had not experienced a cure.

Our study involved an experimental manipulation of the number and balance of anecdotes for each treatment alternative. One group of participants received *balanced anecdotes*, with two testimonials about each treatment—one from a person who got better and one from someone who did not. Another group received *statistically reinforcing anecdotes*: four testimonials from bypass patients, three of whom had gotten better and one who had not (thus, mirroring the 75% success rate of the treatment).

It is important to keep in mind that the anecdotes were uninformative. They did not illustrate anything about the treatment alternatives that people had not already been told. They simply relayed stories of treatment success or failures, and we had already informed them of the success rates of each treatment. Nevertheless, people's hypothetical treatment choices were significantly influenced by the anecdotes they encountered, with 30% of people receiving balanced anecdotes choosing bypass surgery versus 44% of those receiving statistically reinforcing anecdotes. Receiving a larger number of positive anecdotes about bypass surgery increased people's willingness to choose this treatment, even though those anecdotes told people nothing about the treatments that they did not already know.

In a follow-up study, we discovered that we could reduce the influence of anecdotes by providing graphical representations of the success rates of the two treatments (Fagerlin, Wang, and Ubel, 2005). We randomized participants so that half of them received a pictorial representation of the success rates alongside the prose description. (The pictographs that we used are illustrated in figure 20.1.) We found that the influence of anecdotes was eliminated by this pictograph. Regardless of whether we included "balanced anecdotes" or "statistically reinforcing anecdotes," approximately 40% of people chose bypass surgery. In other words, the influence of anecdotes was eliminated when the statistical information was supported by pictographs.

20.1. Pictographs used to communicate cure rate of bypass surgery and balloon angioplasty.

DAs should be judged for invariance. Two DAs that lead to equal comprehension of a specific decision can also lead to different decisions if they introduce any of a number of decisional biases. Therefore, in judging DAs, we should test for such biases, and when they are present we should develop ways to eliminate them.

Correlational Validity

All else equal, a woman with a high risk of breast cancer should be more interested in taking tamoxifen than a woman with a moderate risk, who, on average, should be more interested than someone with a low risk. This is a standard for judging DAs that I refer to as correlational validity. If varying the risk-benefit ratio of a choice has no influence on people's decisions to choose that alternative, then DA developers have to worry that their DA is failing to make the trade-offs clear or is biasing people's choices.

Surprisingly, this standard is not generally used to evaluate health-care DAs. In part, I expect this oversight has occurred because the medical decision-making community has been so firmly wedded to the knowledge model of decision making that they have not felt much need to test whether people are applying their knowledge rationally. In addition, I expect this criterion has been ignored because it does not provide clear guidance about what constitutes a good or bad decision, since there is no way to judge what the appropriate correlation should be between the risk-benefit ratio of an alternative and people's decisions. For example, suppose there is a correlation of 0.1 in breast cancer risk and interest in tamoxifen among women exposed to a given DA. Suppose there is a 0.3 correlation among women exposed to an alternative DA. Is either of these the correct correlation? We might feel confident that a DA that leads to no correlation is flawed, but can we be convinced that one of these DAs is better than the other?

In summary, DAs should be tested for correlational validity, and if the correlation is unacceptably low (which is a judgment call), then the DA should be revised to better highlight the risk-benefit trade-offs.

Time to Process the Decision

In medical practice, it is relatively common for men to receive diagnoses of prostate cancer at the same clinic visit in which a urologist helps them decide how to treat their prostate cancer. More often than not, men leave such visits deciding for or against surgical intervention. Sometimes they choose radiation treatment, but often they have not even had the chance to meet with a radiation oncologist.

There is increasing evidence in the decision-science literature that time is a crucial element of optimal decision making. Although controversial, some studies suggest that unconscious deliberation can improve people's decision making (Dijksterhuis et al., 2006). Such deliberation takes time. There's also ample evidence that people make different decisions when in hot emotional states versus cold (Loewenstein, 1999). With so much information to process and so many options to consider, it hardly seems plausible that a person who has just found out he has cancer would be able to make a good decision quickly.

Thus, DA evaluation should be broadened to include an assessment of whether people had enough time to process the decision.

Adherence

Some decisions are "one and done" affairs—choose surgery over, say, chemotherapy and you will receive the surgery and your decision will be irreversible. But many decisions are not so final. A patient who decides to take a cholesterol pill, for example, faces that decision every day. A person who decides to save more money and reduce entertainment expenditures still faces the temptation to splurge on a nice vacation.

A good DA, then, will not only help people make a decision—about whether to take a pill or a vacation—but will also help them *stick* with the decision. Such DAs should therefore be evaluated for how frequently people adhere to the decisions they make.

Conclusion

I have laid out a few criteria by which to determine whether a structured DA has helped people make decisions that reflect any underlying preferences they have. None of the criteria are on their own sufficient to prove that a DA has led to unbiased decisions. Thus, DA developers need to use careful judgment in applying these criteria to any existing DA, recognizing trade-offs between the potential attributes of a DA and the various outcomes that decision makers care about. When viewed as a whole, these criteria should give decision counselors a much better idea of when they are helping decision makers. These expanded criteria certainly provide a better idea of the strengths and weaknesses of DAs than do the knowledge and satisfaction-based criteria that have dominated the field to date.

References

Baron, J., and Spranca, M. (1997). Protected values. *Organizational Behavior and Human Decision Processes*, *70*(1), 1–16.

Bekker, H., Thornton, J. G., Airey, C. M., Connelly, J. B., Hewison, J., Robinson, M. B., et al. (1999). Informed decision making: An annotated bibliography and systematic review. *Health Technology Assessment*, *3*, 1–156.

Bentham, J. (1907). *An introduction to the principles of morals and legislation*. Oxford: Clarendon Press.

Damschroder, L. J., Zikmund-Fisher, B. J., and Ubel, P. A. (2005). The impact of considering adaptation in health state valuation. *Social Science and Medicine*, *61*(2), 267–277.

Day, R. (2001). Quality of life and tamoxifen in a breast cancer prevention trial: A summary of findings from the NSABP P-1 study. *Annals of the New York Academy of Sciences*, *949*, 143–150.

Diener, E., and Seligman, M.E.P. (2004). Beyond money: Toward an economy of well-being. *Psychological Science in the Public Interest*, *5*(1), 1–31.

Dijksterhuis, A., Bos, M. W., Nordgren, L. F., and van

Baaren, R. B. (2006). On making the right choice: The deliberation-without-attention effect. *Science*, *311*(5763), 1005–1007.

Fagerlin, A., Wang, C., and Ubel, P. (2005). Reducing the influence of anecdotal reasoning on people's health care decisions: Is a picture worth a thousand statistics? *Medical Decision Making*, *25*(4), 398–405.

Gilbert, D., Killingsworth, M., Eyre, R., and Wilson, T. (2009). The surprising power of neighborly advice. *Science*, *323*(5921), 1617–1619.

Gold, M. R., Siegel, J. E., Russell, L. B., and Weinstein, M. (Eds.). (1996). *Cost-effectiveness in health and medicine*. New York: Oxford University Press.

Green, P., and Srinivasan, V. (1978). Conjoint analysis in consumer research: Issues and outlook. *Journal of Consumer Research*, *5*, 103–123.

Griffin, J. (1989). *Well being. Its meaning, measurement, and moral importance*. Oxford: Clarendon.

Holmes-Rovner, M., Kroll, J., Schmitt, N., Rovner, D. R., Breer, M. L., Rothert, M. L., et al. (1996). Patient satisfaction with health care decisions: The satisfaction with decision scale. *Medical Decision Making*, *16*(1), 58–64.

Kahneman, D., Wakker, P. P., and Sarin, R. (1997). Back to Bentham? Explorations of experienced utility. *Quarterly Journal of Economics*, *112*(2), 375–405.

Kennedy, D. M. (2003). On what basis should the effectiveness of decision aids be judged? *Health Expectations*, *6*, 255–268.

Loewenstein, G. (Ed.). (1999). *A visceral account of addiction*. Cambridge: Cambridge University Press.

Loewenstein, G., and Ubel, P. (2008). Hedonic adaptation and the role of decision and experience utility in public policy. *Journal of Public Economics*, *92*(8-9), 1795–1810.

McNeil, B. J., Pauker, S. G., Sox, H. C., Jr., and Tversky, A. (1982). On the elicitation of preferences for alternative therapies. *New England Journal of Medicine*, *306*(21), 1259–1262.

Molenaar, S., Sprangers, M. A., Postma-Schuit, F. C., Rutgers, E. J., Noorlander, J., Hendriks, J., et al. (2000). Feasibility and effects of decision aids. *Medical Decision Making*, *20*(1), 112–127.

Nelson, W. L., Han, P. K., Fagerlin, A., Stefanek, M., and Ubel, P. A. (2007). Rethinking the objective of decision aids: A call for conceptual clarity. *Medical Decision Making*, *27*(5), 609–618.

O'Connor, A. M. (1995). Decisional conflict. In G. K. McFarland and E. A. McFarlane (Eds.), *Nursing care plans: Nursing diagnosis and intervention*. (pp. 468–478). St Louis: Mosby.

———. (1995). Validation of a decisional conflict scale. *Medical Decision Making*, *15*(1), 25–30.

O'Connor, A. M., Bennett, C. L., Stacey, D., Barry, M., Col, N. F., Eden, K. B., et al. (2009). Decision aids for people facing health treatment or screening decisions. Cochrane Database of Systematic Reviews 2003, Issue 1.

O'Connor, A. M., Fiset, V., DeGrasse, C., Graham, I. D., Evans, W., Stacey, D., et al. (1999). Decision aids for patients considering options affecting cancer outcomes: Evidence of efficacy and policy implications. *Journal of the National Cancer Institute. Monographs*, (25), 67–80.

O'Connor, A. M., Jacobsen, M. J., and Stacey, D. (2002). An evidence-based approach to managing women's decisional conflict. *Journal of Obstetric, Gynecologic and Neonatal Nursing*, *31*, 570–581.

Riis, J., Loewenstein, G., Baron, J., Jepson, C., Fagerlin, A., and Ubel, P. A. (2005). Ignorance of hedonic adaptation to hemodialysis: A study using ecological momentary assessment. *Journal of Experimental Psychology: General*, *134*(1), 3–9.

Sen, A. (2004). Capabilities, lists, and public reason: Continuing the conversation. *Feminist Economics*, *10*(3), 77–80.

Smith, D., Loewenstein, G., Jankovich, S., Jepson, C., Feldman, H., and Ubel, P. (2008). Mispredicting and misremembering: Patients with renal failure overestimate improvements in quality of life after a kidney transplant. *Health Psychology*, *27*(5), 653–658.

Smith, D. M., Sherriff, R. L., Damschroder, L., Loewenstein, G., and Ubel, P. A. (2006). Misremembering colostomies? Former patients give lower utility ratings than do current patients. *Health Psychology*, *25*(6), 688–695.

Torrance, G. (1976). Social preferences for health states: An empirical evaluation of three measurement techniques. *Socioeconomic Planning Science*, *10*, 129–136.

Torrance, G. W., Thomas, W. H., and Sackett, D. L. (1972). A utility maximization model for evaluation of health care programs. *Health Services Research*, *7*(2), 118–133.

Tversky, A., and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458.

Ubel, P. A. (2006). *You're stronger than you think: Tapping into the secrets of emotionally resilient people*. New York: McGraw-Hill.

———. (2009). *Free market madness: Why human nature is at odds with economics—And why it matters.* Boston: Harvard Business Press.

Ubel, P. A., Jepson, C., and Baron, J. (2001). The inclusion of patient testimonials in decision aids: Effects on treatment choices. *Medical Decision Making*, *21*(1), 60–68.

Ubel, P. A., and Loewenstein, G. (1997). The role of decision analysis in informed consent: Choosing between intuition and systematicity. *Social Science and Medicine*, *44*(5), 647–656.

Ubel, P. A., Loewenstein, G., and Jepson, C. (2005). Disability and sunshine: Can predictions be improved

by drawing attention to focusing illusions or emotional adaptation? *Journal Experimental Psychology: Applied*, *11*(2), 111–123.

Ubel, P. A., Loewenstein, G., Schwarz, N., and Smith, D. (2005). Misimagining the unimaginable: The disability paradox and healthcare decision making. *Health Psychology*, *24*(4 Supplement), S57–S62.

von Neumann, J., and Morgenstern, O. (1947). *Theory of games and economic behavior* (2nd rev. ed.). Princeton, NJ: Princeton University Press.

Wilson, T., Meyers, J., and Gilbert, D. (2003). How happy was I anyway? A retrospective impact bias. *Social Cognition*, *21*, 407–432.

Woloshin, S., Schwartz, L. M., Moncur, M., Gabriel, S., and Tosteson, A. N. (2001). Assessing values for health: Numeracy matters. *Medical Decision Making*, *21*(5), 382–390.

Zikmund-Fisher, B., Fagerlin, A., Roberts, T., Derry, H., and Ubel, P. (2008). Alternate methods of framing information about medication side effects: Incremental risk versus total risk occurrence. *Journal of Health Communication*, *13*(2), 107–124.

CHAPTER 21

Using Decision Errors to Help People Help Themselves

GEORGE LOEWENSTEIN

LESLIE JOHN

KEVIN G. VOLPP

Many of the most important problems currently facing the United States as well as other developed nations stem from arguably irrational behaviors on the part of individuals. For example, many of the health problems plaguing the United States, such as lung cancer, high blood pressure, and diabetes, are so-called lifestyle diseases that are exacerbated by unhealthy behaviors (Schroeder, 2007). Modifiable behaviors such as tobacco use, obesity-related behaviors, and alcohol abuse account for nearly one-third of all deaths in the United States, which only spends 2%–3% of the $2.1 trillion spent on health each year on prevention (Flegal et al., 2005; Mokdad et al., 2000; Satcher, 2006; Woolf, 2007). Moreover, although there is an expanding array of beneficial medications available to deal with these and other health conditions—for example, to control blood pressure and cholesterol and to avoid strokes—the benefits of these medications are far from fully realized, in large part due to poor adherence rates among patients. Thus, for example, about half of patients who have a heart attack had stopped taking their cholesterol medication within a year of their heart attack (Jackevicius, Mamdani, and Tu, 2002). Likewise, as we discuss below, adherence to medication that prevents strokes, which is extremely inexpensive and effective, is remarkably low even in the best circumstances—a clinic devoted to administering it (Chiquette, Amato, and Bussey, 1998).

Other problems show similar patterns. For example, prior to the economic downturn, the savings rate in the United States was –1%, that is, individuals, on average, spent more than they earned. In 2000, the median net worth of American households, excluding home equity was $13,473, and for households of age 65 and older, not much more—$23,369 (U.S. Census Bureau, 2003, table A). Currently, only 40% of Americans put money aside in company-sponsored 401(k) plans, and the median American family owns no stocks, even in retirement accounts (Bucks, Kennickell, and Moore, 2006). Yet, the average American family spends a staggering $4,000 per year on gambling (ignoring the amount they receive back in the form of winnings). In a survey commissioned by the consumer federation of America in 2003, 86% of people said that financial planning was important to them, but only 46% indicated that they had developed such a plan (Consumer Federation of America, 2003). Americans want to save money, but many are failing to do so.

Standard Economics and Behavioral Economics

Economics is the discipline that is most closely associated with public policy. However, traditional economics is ill-equipped to deal with problems arising from suboptimal individual behavior because it is built on a rational-choice paradigm that effectively assumes that such problems do not exist. Thus, prominent economists have argued that addiction is the outcome of a rational choice (Becker and Murphy, 1988), that people are obese because they have judged that the pleasure of eating is worth the discounted costs (Murphy, 2006), and that suicide is a rational choice for those who judge that "the total discounted lifetime utility . . . reaches zero" (Hamermesh and Soss, 1974). The implication of analyses such as these is that interventions to reduce addiction, obesity, or suicide are likely to be counterproductive, since those who choose these behaviors are making an optimal decision to do so.

Behavioral economics is in a better position to provide policy solutions to problems that arise from individual behavior because it not only acknowledges that behavior is often far from optimal, but also identifies a variety of decision errors and judgmental biases that contribute to such departures from optimality. The central point of this paper is to argue that many of the same decision errors that produce self-destructive behavior can be used to people's benefit.

Behavioral Economics and Asymmetric Paternalism

By recognizing that even mature adults are subject to systematic decision errors, behavioral economics provides a potential rationale for paternalistic policies—policies intended to help individuals by improving the choices they make. Much like parents who intervene in the diets of their children, based on the assumption that children often do not know what is best for themselves and, even when they do, often do not act on that knowledge, behavioral economics opens the door to analogous policies applied to adults.

However, paternalism in its traditional, "heavy-handed," form elicits widespread discomfort, and for good reason. One concern is that policy makers may not know what is best for individuals (see, e.g., Glaeser, 2006); a related argument is that people may have good reasons (that policy makers do not recognize) for behaving as they do. There is also a fear of regulatory capture, whereby paternalistic interventions ostensibly intended to protect individuals are in fact designed to help those being regulated. For example, it has been argued that cigarette companies knew that warning labels on cigarette packages would not deter smoking (the apparent intention of labels) but hoped that such labels would help to shield them from liability for health damages. Finally, by reducing, or even removing, individuals' freedom of choice, heavy-handed paternalism is unappealing to the many people, including many behavioral economists, who believe that autonomy of choice has inherent value.

Asymmetric paternalism (Camerer et al., 2003) seeks to obtain some of the benefits of paternalism while skirting the pitfalls of heavy-handed paternalism. It is based on two central tenets. First, paternalism is unavoidable: behavior is shaped by people's environments, and environments have to be structured in some way; there is no neutral way to structure an environment. Thaler and Sunstein (2003) illustrated this point with the example of a manager at a company cafeteria who is concerned about the well-being of employees and recognizes that they tend to load up on the first food they encounter in the food line. Deciding how to order an assortment of food in a cafeteria is unavoidable; food must be ordered *somehow*. Given that inevitability, Thaler and Sunstein ask, why should the cafeteria manager not arrange the healthy food first in line so as to promote employee health?

The second tenet of asymmetric paternalism, which is also nicely illustrated by the cafeteria line, is that it is often possible to design interventions that help those who are behaving suboptimally without restricting the choices of those behaving optimally. In the case of the food line, people who mindlessly load up on the first food they encounter will eat more healthily, but someone who knowingly craves the double-cheese lasagna is at complete liberty to indulge that preference (For more on how people can be "nudged" to make better choices, see Thaler, Sunstein, and Balz, this volume).

There is, of course, some risk that overly zealous paternalists will go crazy engineering environments that direct people to conform to their own idiosyncratic views of what is best. However, we suspect that most asymmetric paternalistic interventions will be aimed at more prosaic goals that most people would embrace: quitting smoking, losing weight, saving for retirement, etc. In this chapter, we will skirt the metaquestion of what it means for a person's behavior to be in their "best interest." We simply assume that if an individual expresses a consistent desire to achieve a certain goal, such as losing weight, taking medications or saving money, it is relatively unobjectionable to help them achieve that goal in a fashion that does not restrict their ultimate freedom of choice.

Using Decision Errors to Help People: The Theory of the Second Best

The "theory of the second best" (Lipsey and Lancaster, 1956) refers to the situation that arises when one or more of the conditions for economic optimality are not satisfied. If one of the conditions for economic optimality is not satisfied, the theorem states, it is possible that economic efficiency will be best served by deviating from other conditions for optimality. That is, the second-best solution in a situation in which the first-best solution is not possible may involve other deviations from the conditions that are usually assumed to be optimal.

Although the theory of the second best was originally applied to market-level phenomena—to deviations from the characteristics of a "perfect market"—the same logic applies at the individual level. If an individual's behavior deviates from optimality in one way—for example, as a result of being excessively averse to taking risks, of overweighting immediate gratifications, or of being overconfident—the individual's best interests may not be served by behaving optimally in every other respect. Expressed more intuitively, decision errors can cancel one another out. Individuals will

not do as well making two errors as they would if they made no errors, but if those errors counteract one another sufficiently, people may do better making two errors than they would if they made only one error.

An example of such an error-canceling situation (although not cast by the authors in terms of the theory of the second best) was discussed by Kahneman and Lovallo (1993) in the context of entrepreneurship. They point out that if potential entrepreneurs are excessively averse to taking risks due to a distaste for experiencing losses (a phenomenon known as "loss aversion"), then it may actually be beneficial for them to also be overoptimistic about their chances of success. Entrepreneurs would do best to not be loss averse *or* overoptimistic, but if these errors balance one another out to some extent, they might do better on average if they exhibit both errors than if they exhibit only one.[1]

Specific Decision Errors and How They Can Be Used to Improve Behaviors

Most, although not all, decision errors can be interpreted as instances of *misweighting*—of putting either too much weight or too little weight on specific types of costs and benefits. Although such misweighting generally degrades the quality of decision making when it occurs and is unavoidable, as suggested in the last section of the paper, it is sometimes possible to use other decision errors to produce a kind of compensatory *reweighting* that offsets the initial misweighting.

The Status-Quo, or "Default," Bias

The status-quo, or default, bias (e.g. Johnson and Goldstein, 2003; Kahneman, Knetsch, and Thaler, 1991; Samuelson and Zeckhauser, 1988) refers to people's tendency to take the "path of least resistance"—to keep doing what they have been doing, or to do what comes automatically, even when superior alternatives exist. Defaults have been blamed for a wide range of suboptimal outcomes, from the failure of employees to put aside retirement funds in companies with a default contribution rate of zero (Gneezy and Potters, 1997; Madrian and Shea, 2000; Thaler and Benartzi, 2004), to suboptimal allocations between investment alternatives (Thaler et al., 1997), to excessive ingestion of fries and large sodas as part of "supersized" meals at McDonald's (Halpern, Ubel, and Asch, 2007; Loewenstein, Brennan, and Volpp, 2007; Thaler and Sunstein, 2003). However, as many behavioral economists have pointed out, defaults (see, e.g., Johnson and Goldstein, this volume), if chosen judiciously, can also be used to propel people toward self-beneficial behaviors. Thus, if individuals tended to be pathologically risk averse, by making the default investment portfolio riskier than people would naturally choose, people could be steered in the direction of incurring a more optimal level of risk.

Loss Aversion

Loss aversion is the tendency for people to put substantially greater weight on losses than gains (e.g. Kahneman, Knetsch, and Thaler, 1991; Tversky and Kahneman, 1991). It can produce a variety of suboptimal patterns of behavior, from pathological risk-aversion (as already discussed) to the tendency for people to hold on too long to houses (Genosove and Mayer, 2001) or stocks (e.g., Odean, 1998; Shefrin and Statman, 1985; Weber and Camerer, 1998). However, the same property that makes loss aversion destructive in some situations—its tendency to amplify the weight put on specific outcomes if they are framed as losses—can be used to advantage when people's natural tendency is to underweight outcomes. Thus, for example, if people are putting too little weight on delayed outcomes because they discount the future excessively, framing those delayed outcomes as losses can potentially increase the weight put on them—correcting one error with another.

Present-Biased Preferences

Present-biased preferences (e.g. Ainslie, 1975; Frederick, Loewenstein, and O'Donoghue, 2002; Loewenstein, 1992; Loewenstein and Angner, 2003; O'Donoghue, and Rabin, 1999, 2000), which are also referred to as hyperbolic time discounting, encompass two important behavioral propensities: (1) the tendency to overweight immediate costs and benefits relative to those occurring at any point in the future, and (2) the tendency to take a much more evenhanded approach to delayed costs and benefits occurring at different points in time. The notorious resolutions that one will begin to diet or save money *tomorrow* encompass both propensities: the overweighting of immediate costs deters one from the immediate misery of dieting or saving, but the more evenhanded perspective on future time makes one willing to impose these costs on oneself in the future. These two properties, in turn, suggest two ways that present-biased preferences can be used to advantage. First, the tendency to overweight immediate costs and benefits suggests that the motivational impact of costs and benefits—for example, rewards for good behavior or punishments for bad behavior—can be greatly increased by making them immediate, ideally

coinciding as closely as possible with the timing of behaviors they are attempting to encourage or deter. Second, the more even-handed attitude toward different times in the future suggests that people will be willing to commit to self-control devices that will be activated in the future that they would not be willing to commit to if they were to be activated immediately.

The Self-Serving Fairness Bias

The self-serving bias (e.g. Babcock et al., 1995) is the tendency for people to confuse what is in their own interest with what is fair. One of the hallmarks of the self-serving bias is the belief that one's biased view is in fact not biased but rather an impartial representation of reality—a phenomenon that Ross and Ward (1996) label naive realism. The upshot is that to the extent that parties believe that their own perspective reflects reality, they are also likely to think that their perspective will be shared by a neutral third party.

For example, the self-serving bias has been shown to play a critical role in negotiation impasse—in the failure to achieve settlement of a dispute even if it is in all parties' interests to do so. If people's perceptions of fairness are biased in a self-serving fashion, there may be no possible negotiated solution that all parties would perceive as fair. Again, however, this bias can be used to advantage in some situations by exploiting the fact that people tend to believe that their own biased perspective is neutral and objective and, hence, that it will be shared by a neutral third party. If people are convinced that a neutral third party will share their perspective, they may be willing to settle a dispute via arbitration, assuming an arbitrator can be located who, prior to rendering a decision, is perceived by all sides to be unbiased. Therefore, this bias, and in particular, people's ignorance that they are subject to it, can be used advantageously. Other chapters suggest ways in which this bias may potentially be overcome (Pronin and Schmidt, this volume; Ross, this volume).

Nonlinear Probability Weighting

Nonlinear probability weighting is another two-part effect (like present-biased preferences) that can be exploited to advantage. As encompassed in prospect theory (Kahneman and Tversky, 1979), (1) people tend to put disproportionate weight on outcomes that have a small probability of occurring but (2) also tend to be insensitive to variations in probability at the low end of the probability scale, a pattern Sunstein (this volume) refers to as probability neglect. Because people draw little distinction between, for example, a .00001 versus .0000001 chance of winning a prize, even though the probabilities differ by several orders of magnitude, such overweighting is especially extreme for very small probabilities. The overweighting of small probabilities has various negative effects on decision making, such as, undoubtedly, contributing to the popularity of lotteries. However, the overweighting of small probabilities can be exploited by giving people lottery prizes rather than fixed amounts of money for good behavior, providing more "bang for the buck" from economic incentives that are designed to help people to engage in beneficial behaviors.

Peanuts Effects

The peanuts effect (Markowitz, 1952; Prelec and Loewenstein, 1991; Weber and Chapman 2005) is the common tendency to put little weight on very small outcomes—both gains and losses.[2] Like the overweighting of small probabilities, the peanuts effect encourages lottery play because the $1 cost of a lottery ticket is viewed as peanuts – as "chump change" (Haisley, Mostafa, and Loewenstein, 2008). Generalized somewhat, the same term can encompass the underweighting of nebulous or amorphous, often delayed, consequences, which can help to explain such diverse self-destructive patterns of behavior as snacking, cigarette smoking, and talking on the cell phone while driving. In each of these cases, the benefits of the activity—for example, the pleasure of eating or smoking a cigarette—are immediate and tangible, but the costs—an infinitesimally small increase in the chance of lung cancer and other diseases, an imperceptible increase in weight, or a tiny increase in risk of injury or death—are amorphous. The peanuts effect can be viewed as a form of underweighting; however, again, this decision error can in some situations be channeled to help people rather than to hurt them. For example, the same tendency to underweight small outcomes that leads one to eat "one more chip" over and over can also make it relatively painless for people who might have trouble saving a large chunk of money in one fell swoop to instead make a large number of much smaller deposits, each one of which feels relatively painless.

Narrow Bracketing

Choice bracketing (e.g. Read, Loewenstein, and Rabin, 1999) is the process of grouping individual choices together into sets. When making choices, people can either group them *broadly*, by considering all of the consequences taken together (as standard economic theory assumes) or *narrowly*, by making each decision in isolation. A *bracketing effect* occurs when choice

outcomes under narrow bracketing differ from those under broad bracketing, and a general finding is that people tend to bracket narrowly: they myopically focus on the local consequence of the most immediately available choices and ignore the aggregated costs and benefits over a long time horizon (e.g., Herrnstein and Prelec, 1992; Sabini and Silver, 1982). This tendency becomes especially pronounced in temporal bracketing contexts where choices are made sequentially, the classic example being a phenomenon known as myopic loss aversion, in which risk aversion is heightened to the extent that investment decisions are made one decision at a time, neglecting the consequences of aggregation (Benartzi and Thaler, 1995; Gneezy and Potters, 1997). Bracketing effects interact with many other biases and can be used as a tool to induce these other biases. For example, the peanuts effect is more likely to occur when costs or benefits are framed narrowly, so, to the extent that the peanuts effect can be used to help people help themselves, narrow bracketing can in turn be used to increase the likelihood that people frame outcomes as "peanuts."

Projection Bias and Hot-Cold Empathy Gaps

Projection bias (Loewenstein, O'Donoghue, and Rabin, 2003) is the tendency for people to project their current preferences onto the future. Hot-cold empathy gaps, which often underlie projection bias, are the tendency for people to underestimate the impact of current emotions and drives and to fail to predict the impact of such emotions and drives on their own future behavior. People who are not hungry, for example, mispredict their own future food choices, overestimating the likelihood that they will choose healthy options (Read and van Leeuwen, 1998) and judge other people who fail to show dietary moderation more harshly than they do when they themselves are hungry (Nordgren, van der Pligt, and van Harrevald, 2008).

Projection bias leads to diverse suboptimal patterns of behavior—from over-shopping on an empty stomach to excessive seeking of wealth and status (because one fails to anticipate the extent to which one will adapt to either). However, as we show below, because projection bias can cause people to underappreciate the misery of future self-denial, it can be used to encourage people to precommit to self-binding measures that help them to accomplish their long-term goals.

Overoptimism

Self-predictions of future behavior are systematically biased toward being overly optimistic. For example, research on the planning fallacy has shown that individuals tend to underestimate their task completion times (Buehler, Griffin, and Ross 1994, 2002). In the moral sphere, people have been found to overestimate the likelihood that they will engage in prosocial behavior, such as donating to charity (Epley and Dunning, 2000). In the health domain, people overestimate their future gym usage, and as a result opt for paying a flat rate for gym memberships, even though most would spend less if they were to pay on a per-visit basis (Della Vigna and Malmendier, 2006). Mail-in product rebates are a frequently cited example within the marketplace. Although such rebates have been shown to promote sales, only a small number of rebate coupons (5%–20%) are typically redeemed (Bulkeley, 1998). The optimistic bias apparent in people's self-predictions of their future behavior is especially striking given that in each of these examples the target behavior is largely under the individual's control. Later, we show how, when combined with projection bias, overoptimism can be used to facilitate weight loss.

In conclusion, as summarized in table 21.1, a wide range of biases that normally detract from the quality of decision making can be exploited in policies designed to enhance beneficial behaviors. The next section reviews a variety of such initiatives, including some that have been already tested and others that are still in the design phase.

Applications at the Individual Level

Saving

Perhaps the single most significant application of behavioral economics to public policy, so far, has been saving behavior. The problem of undersaving in the United States is particularly concerning because, far from implementing the types of policies that would be suggested by behavioral economics, the United States has been moving in the opposite direction. The easiest way for people to save is to have it done automatically, without the need for decision making or the imposition of self-control. The defined benefit pension plans that used to be the norm for moderate- and large-scale employers did just that; they required little if any decision making or deliberate deferral of gratification on the part of employees. The pervasive shift from defined benefit to defined contribution savings plans, however, shifted the burden of decision making and of deferral of gratification to the employee. In defined benefit savings plans, individuals have to save for their own retirement but get a tax break from the government as well, often as help from their employer in the form of a match on savings. Hence, the theory of the second best comes into play again, although in

Table 21.1 Biases that can be exploited in policies designed to enhance beneficial behaviors

Bias	How it can be used to a person's advantage
Status-quo/default bias	Make options that reflect a "correct" weighting of costs and benefit the default.
Loss aversion	Frame underweighted outcomes as losses; overweighted outcomes as (forgone) gains.
Present-biased preferences	Make rewards for beneficial behavior frequent and immediate. More evenhanded approach to delayed costs and benefits. Get people to commit to self-interested behavior ahead of time.
Self-serving fairness bias	Can be used to promote dispute-resolution (because negotiators underestimate the likelihood of judgments they view as unfair).
Nonlinear probability weighting	Provide probabilistic rewards for self-interested behavior.
Peanuts effect	Focus on small but frequent behaviors to increase the tangibility of underweighted costs and benefits, and decrease the tangibility of overweighted costs and benefits.
Bracketing	Bracket behavior narrowly. For added potency, combine with other decision errors (e.g. loss aversion; peanuts effect).
Projection bias and hot-cold empathy gaps	Set up mechanisms through which binding self-commitments are made in "cold" states.
Overoptimism	Use overoptimism to encourage precommitment.

a somewhat different way from that described earlier. The ideal—the first best—would have been to continue with defined benefit plans, albeit perhaps with modifications to enhance portability and to ensure the solvency of the underlying funds. However, given that we are not in this first-best situation, the fallback is to use ideas from psychology, specifically to exploit decision errors, to help ensure that people save for their retirement.

The main policy response to concerns about shortfalls in saving has been the usual economic remedy—to increase the effective return on saving by offering various tax breaks on defined contribution plans. There are, however, several problems with such an approach. First, it assumes that people are making a rational, deliberate trade-off between current and future consumption, but judging from the weak relationship between natural variations in interest rates and savings rates, the problem of undersaving is not mainly due to the perception that returns on saving are too low. In fact, standard economic theory is largely silent about the impact of an increase in the rate of return on saving on savings rates, given that a change in returns produces both a substitution effect (which makes future consumption more attractive) and an income effect (which renders saving for the future less necessary). Second, an increase in effective returns induced by tax exemptions is extremely inequitable, because the benefits accrue disproportionately to those in the highest tax brackets, and inefficient, because those in the highest tax brackets who get the biggest tax discounts are already those who are most likely to save adequately; the problem of undersaving is a much bigger problem for low- and lower-middle-income families.[3]

Unlike approaches based on conventional economics, the essence of all interventions proposed and tested by behavioral economists has been to make increased saving the path of least resistance. Unlike attempts to increase saving through tax breaks, which result in a loss of tax revenue and yield no benefit to the extent that the money would have been saved even if the tax breaks were not offered, the behavioral remedies do not require additional government outlays or reductions in tax collections.

DEFAULTS

The best-known interventions to increase savings have involved changing default contribution levels to 401(k) plans (see discussion of the default bias above). For example, Madrian and Shea (2000) studied a company that changed from a default employee contribution rate of 0% to 3% and observed a steep increase in the fraction of employees saving through the plan, as well as an increase in average contribution rates. However, the change was not without problems. The company offered a 6% match; employee contributions

were matched one-to-one by the employer up to 6%, so the optimal level of contribution from the perspective of the employee was 6% (see Choi, Laibson, and Madrian, 2005). However, the percent of employees contributing 6% actually dropped after the plan was implemented, and some employees who would have saved at 6% instead saved at 3%, reflecting the power, but also the potential pitfalls of defaults. Moreover, the default investment allocation was to 100% money market, and, again reflecting the power of defaults, most employees left this allocation unchanged, whereas a much higher percentage had invested in stock before they got defaulted into the money market. This intervention and others like it are discussed in detail in other chapters of this volume (Benartzi, Peleg, and Thaler; Thaler, Sunstein, and Balz).

SAVE MORE TOMORROW

A very clever and highly successful program to increase savings devised by Thaler and Benartzi (2004) provides perhaps the single best example of using errors to help people. In their program, employees precommit to diverting some fraction of future wage increases into a retirement account. For example, an employee who could anticipate at least a 4% yearly increase in salary over upcoming years could elect to have half of that increase put into a retirement account over the next several years. Save More Tomorrow (SMarT) plays on three different biases. First, the save more *tomorrow* feature plays on the structure of present-biased preferences, and specifically on people's willingness to make far-sighted decisions for the future as long as they do not entail immediate sacrifice. Second, the fact that increments in saving come out of future wage increases plays on the idea that forgone gains are far less painful than out-of-pocket losses (Thaler 1980, 1985). Finally, the SMarT plan takes advantage of the status quo/default biases: without the human tendency to inertia, it is likely that people would change their mind about saving the money once tomorrow became today.

OTHER POSSIBLE APPROACHES

Currently, Emily Haisley and George Loewenstein are working on two programs to promote savings using lottery inducements. One program involves the design of a completely new type of state lottery ticket that allows customers to *simultaneously* play the lottery and save money. The proposed program draws on the same biases that make playing the lottery so attractive to people. A portion of the ticket's price is wagered in a typical lottery game and the remainder is deposited into a savings account. These tickets would be sold through automatic ticket vending machines that also track account balances. An important feature of the program is an added incentive to continue to save and to keep account balances high. Each month, savers automatically get one bonus ticket for every $100 they have on deposit in their account, which gives them the chance to win additional cash prizes. In addition, they receive a communication every time they add another $100 to their account.

This program is designed to help low-income individuals overcome procrastination to save. Saving is challenging in part due to time discounting: the costs are immediate but the rewards are delayed far into the future. The peanuts effect also contributes to difficulty saving because any act of abstention from spending is likely to have a minimal impact on savings. This program plays on present-biased preferences and the overweighting of small probabilities by providing an *immediate* probabilistic reward for saving. Beside providing a motivation for saving, the pleasure and entertainment value of playing the lottery helps to negate the pain of self-denial.

The program also plays on the peanuts effect. People may dislike the pain of setting aside large sums of money all at once, but this program enables individuals to make small, frequent deposits. The program is likely to be particularly effective for low-income individuals, who, in addition to feeling this pain of saving, may have so little economic slack (as Mullainathan and Shafir call it in this volume) that they are unable to make large deposits. In addition, the ubiquity of lottery sales kiosks provides frequent reminders to purchase tickets.

A final feature of the program plays on the differential weighting of opportunity costs (foregone gains) and out-of-pocket costs (as discussed above in connection with the SMarT plan). The lottery is set up to give a high probability of a relatively small prize (e.g., $30) and a very small probability of a very large jackpot. Although savers are informed of their winnings so they can fully enjoy their good fortune, the smaller winning amounts are, in fact, automatically deposited into the individual's savings account, reducing the temptation to spend.

Whether such a program would be beneficial depends on who, if anyone, would end up purchasing the new type of lottery ticket and where the money to make the purchases would come from. Ideally, purchases would be concentrated among people who are already playing the lottery and who would switch from purchasing conventional lottery tickets to purchasing the savings tickets. Much less ideal would be if the new lottery tickets brought people in to playing the lottery who were not playing before, and worse, if it led them to play the lottery with money

they otherwise would have put into saving. Clearly, a small-scale market test of such tickets would be desirable before they were introduced on a grand scale.

The second program involves an innovative design for individual development accounts (IDAs). IDAs are matched savings accounts for low-income individuals that are typically geared toward purchasing a home, paying for education, or starting a small business. IDAs usually employ a 2:1 match rate that allows the account holder to withdraw $2 for every $1 deposited, but only after reaching the savings goal. The same goal of encouraging saving can potentially be achieved at a much lower cost by replacing the guaranteed match with a lottery incentive. In the specific program being tested (Loibl, Haisley, and Loewenstein, in preparation), savers are guaranteed a fixed match of 1:1 on any money they put aside, and, in addition, are offered a lottery match. Specifically, there is a 1 in 10 chance that any amount they put aside will be matched 5-fold, and a 1 in 50 chance that any amount they put aside will be matched 25-fold.

Although tests of this idea are ongoing, there is a wealth of evidence that lottery-linked savings accounts can be applied successfully in low-income populations. In contrast to the IDA program just outlined, which offers a probabilistic match on deposits, most lottery-linked accounts offer prizes that are connected to balances rather than deposits. For example, many commercial banks outside of the United States offer lottery-linked savings accounts in which monthly drawings are held for cash and prizes, and customers get one lottery ticket for every $\$X$ they have on deposit for the duration of the month (Guillen and Tschoegl, 2002). Similarly, many governments issue "prize" bonds, which periodically distribute the interest to just a few bond holders, and microfinance institutions give depositors "saving cards" that offer the chance to win prizes if a lottery drawing matches a portion of the serial number on the card. All of these lottery-linked accounts have been shown to draw customers from the lower end of the income distributions (see, e.g., Tufano, 2008). They benefit such customers by increasing their financial security, although invariably they offer reduced (and often zero) interest rates, with the difference used to cover the costs of the prizes.

Improving Health Behaviors

Schroeder (2007) highlighted the poor state of health outcomes in the United States relative to those in other developed countries and pointed out that the greatest opportunities for improvement in health do not involve further improvements in health-care delivery but, rather, changes in individual health behaviors. Schroeder also notes that obesity and smoking, despite the reductions in prevalence of smoking over the past several decades, are the two most significant contributors, with smoking contributing to more than 400,000 deaths per year in the United States.

Whether these potential improvements in health can be achieved, however, depends on whether it is possible to change health behaviors. Clearly, the answer does not lie in the standard economic prescription—that is, providing more information. People are acutely aware of the health hazards of smoking. Indeed it has been argued that smokers tend to overestimate these hazards (Viscusi, 1992; although there is controversy on the issue; see Slovic, 2001), in which case giving people better information might only cause them to smoke *more*. Furthermore, about 70% of smokers say they want to quit smoking although only about 2%–3% per year succeed (Bartlett et al., 1994; Hughes, 2003). The problem is probably not the result of poorly informed decision making but rather of being unable to implement good intentions.

WEIGHT LOSS

Losing weight seems to be one of the most difficult goals to accomplish. In our hyper-weight-conscious society, people are powerfully motivated to lose weight yet are mostly unable to do so. The problem is so seemingly intractable that one prominent diet researcher, Janet Polivy, has coined the term *false hope syndrome* to describe the unfounded optimism of those who attempt different weight-loss strategies. The same researcher has conducted clinical tests of what she labels the "undiet," which simply involves giving up on the false hope of dieting. In one study comparing the undiet to various more optimistic dieting strategies, Polivy found that dieters and undieters gained about the same amount of weight, but those on the undiet reported fewer neurotic patterns of behavior and lower levels of depression (Polivy and Herman, 1992). If conventional diets do not work, does behavioral economics have any insights to offer about what might?

Results of a three-arm randomized controlled weight-loss trial (Volpp et al., 2008) provide hope that ideas from behavioral economics can be productively applied to weight loss. The study used financial incentives to motivate weight loss—loss aversion, overoptimism, and regret aversion—to help overweight people lose weight. Study participants (veterans in Philadelphia) were enrolled in a weight-loss program the goal of which was to lose 16 pounds in 16 weeks.

Two different types of incentive conditions were used and compared to a no-incentive control: a

lottery-based incentive and a deposit contract incentive. Study participants in the incentive conditions were required to call in their weight to the study nurse each day and were given daily feedback via text pagers. Accumulated incentives were paid out on a monthly basis once phoned-in weights were confirmed by a monthly weigh-in that took place at the clinic. This strategy played on loss aversion, because winnings during the month were received only if the participants continued to lose weight throughout the month and were below the monthly goal at the end-of-the-month in-person weighing. The combination of daily feedback but monthly payments has several advantages: (1) playing on present-biased preferences, and specifically the overweighting of immediate benefits, it gives people who attain their goals frequent positive feedback in the form of messages that they have been paid; (2) however, by paying people only monthly, it increases the likelihood that a significant amount of money will have been accumulated, thus avoiding potential peanuts effects; (3) finally, by giving both symbolic rewards delivered by message *and* real rewards delivered in the form of an immediately cashable check, it leverages the payments maximally; it is almost as if each payment is made twice.

The lottery incentive condition consisted of a daily lottery with an expected value of $3 per day (1 in 5 chance of winning $10, 1 in 100 chance of winning $100), with subjects eligible for payment each day if they were on track to achieve their monthly weight-loss target. The design was motivated by the idea that lotteries tend to have greater incentive value than certain payments of the same expected value (see overweighting of small probabilities, discussed in "Nonlinear Probability Weighting," above), and that lottery players are motivated by both a forward-looking element (deriving from anticipation of the large payoff) and a backward-looking element based on the frequency of wins in the recent past (Camerer and Ho, 1999). Subjects were informed daily of the lottery outcome via their text pagers.

The lottery incentive condition also capitalizes on *regret aversion* by informing subjects who failed to attain their daily goal of whether they *would have* won had they met their target weight that day. Like the IDA savings program already discussed, the lottery intervention also plays on present-biased preferences by giving subjects rapid positive feedback for beneficial behaviors.

In the second incentive condition, deposit contract, subjects could deposit $.01–$3.00 per day of their own money, which was matched 1:1. Subjects reported their weight daily and received the sum of both amounts each day that they were on track to meeting their monthly weight-loss targets, but they forfeited their deposit and match if they were not. They also received a fixed payment of $3.00 each day they were under their targets.

The deposit contract condition plays on subjects' overoptimistic self-predictions (see the discussion in "Overoptimism," above). People tend to be overly optimistic in predicting how much weight they will lose (or similarly, fail to appreciate how difficult it is to lose weight); therefore, when asked to put money down at the beginning of the month toward attaining their weight-loss goals, about 91% of subjects were willing to do so, and of these participants, the average deposit contract increased during each month of participation, from $1.35 in month 1 to $1.59 in month 2 to $1.83 in month 3, leveling off to $1.85 in month 4. As the subjects struggled with losing weight, their desire to avoid losing the deposit provided added motivation to attain the weight-loss goal. Bound by their optimistic predictions and averse to losing their deposits, these participants ideally had their biases turn into a self-fulfilling prophecy.

The results of both interventions were dramatic: incentive participants lost over three times more weight than the controls. Whereas lottery and deposit contract participants lost an average of 13.1 and 14.0 pounds, respectively, the mean weight loss was significantly lower in the control condition (M = 4.0 pounds; Volpp et al., 2008a).

The appeal of this approach was also supported by the extremely low drop-out rate in the study. Only 9% of subjects dropped out of the study, a lost-to-follow-up rate that was much lower than is typical in weight-loss intervention studies, where rates are often as high as 40%–50%. Among subjects not lost to follow-up across both incentive arms, participants called in daily weights more than 90% of the time, indicating the feasibility of an approach that probably keeps weight loss salient among participants. The study's impressive results attest to the power of applying principles from behavioral economics to promote health behavior.

In a related vein, Wansink (this volume) discusses policies that could lead people to make healthier choices effortlessly. As in this chapter, where we argue that decision errors can be used to offset one another, the central premise of Wansink's discussion is that eating cues, such as packaging size, can be reversed to help people eat less food rather than more.

Despite the success of our weight-loss study, there are several caveats that must be acknowledged. First, once the incentives were removed at the end of the four-month study period, the participants in the two treatment groups gained back a significant fraction of the weight they had lost. Currently, we are testing whether this outcome could be avoided by running

a study in which incentives are offered for a longer period of time. Second, the program was relatively expensive and complicated to administer. Although we are currently testing the cost-effectiveness of removing the $3 fixed payment from the deposit contract incentive, beyond this fixed payment, there are still significant costs, such as staffing the clinic for the monthly weigh-ins, processing phoned-in weights, and sending out text messages. If these functions could be automated (which now seems possible using available technology), the costs of running a program using deposit contracts would be substantially lower.

FURTHER (UNTESTED) WEIGHT-LOSS APPLICATIONS

Framing and Bracketing in Weight-Loss Programs

As outlined in the section on specific decision errors, people tend to bracket decisions narrowly and to be susceptible to framing effects, yet these phenomena could be combined to people's advantage to facilitate weight loss. It is conceivable that the benefit of framing a diet broadly or narrowly might depend on a person's stage of dieting. Framing a weight-loss program broadly may make people particularly likely to sign up for one; indeed, the advertisements for many weightloss programs aggregate the amount of weight to be lost over the course of several months (e.g., "lose 10 pounds in 2 months" as opposed to "lose 0.16 pounds a day for 2 months"). Such a frame emphasizes the total weight loss while simultaneously downplaying the daily "grunt work" necessary to lose the weight. In other words, the broad frame may give an illusion of losing weight with minimal effort, thus helping to motivate people to initiate such a diet.

During intermediate stages of a diet, however, switching to narrow framing may make weight loss more manageable by breaking down the overall goal into subgoals that are easier to attain and monitor. This line of thinking is consistent with Gollwitzer's (1999) notion of implementation intentions. Applied to our weight-loss study, which we described earlier, the benefits of narrow framing could help to account for the success of our intervention. Participants in our program were required to monitor their weight on a daily basis (whereas controls were not). Further research will hopefully disentangle the effects of incentives and feedback on weight loss.

Finally, it may be helpful to switch back to broad framing toward the end stages of a diet, as a person approaches his goal, because doing so highlights the impressiveness of the overall weight loss. Because of goal gradients (Kivetz, Urminsky, and Zheng, 2006), such an emphasis is likely to be particularly motivating at the end stages of the diet.

Stimulating People to Exercise

Beside dieting, of course, the other route to weight loss is exercise. Beyond weight loss, exercise has myriad benefits for physical and mental health, and even for cognitive functioning (e.g., Colcombe and Kramer, 2003; Folkins and Sime, 1981). Is it possible to use decision errors to encourage people to exercise more?

To some degree, decision errors already work to prompt people to exercise. People find flat-rate gym payment plans more palatable than per-visit ones, despite the fact that based on their usage (or rather, lack thereof), they would spend less if they were to pay on a per-visit basis (Della Vigna and Malmendier, 2006). This is sometimes referred to as the flat-rate bias (see, e.g., Lambrecht and Skiera, 2006). The flat-rate bias favors exercising because, after having joined an exercise club based on attraction to the flat rate, people are then often motivated to get their money's worth. Thus, the attempts by one author's mother to "get the price of a run down to $2" on family ski trips. Such a tendency could be further amplified by giving people each time they visit the gym their "new per-visit price," which would decline the more they visited.

One can imagine, however, schemes that would go even further toward encouraging gym usage. For example, customers could be offered discounted flat-rate memberships if they pledge to visit the gym a certain number of times per month and agree to pay a fine if they do not attain this quota. Requiring fines to be paid in cash would make them particularly "painful" (Prelec and Loewenstein, 1998), thus making them an even more powerful detractor of underusage. Similar to the weight-loss deposit contract plan already discussed, such a scheme could make overoptimism self-fulfilling. Similar to people's overoptimism about adhering to a diet, people are likely to be overly optimistic about their propensity to exercise, leading them to be willing to accept fines if they do not exercise at a fairly high rate. Once they have implemented the fine scheme, however, people will be motivated by loss aversion to avoid being fined. This scheme is designed to stimulate exercise without limiting freedom of choice: customers are free to choose the higher-priced plan that does not require minimum monthly usage. Moreover, although we suspect this plan would generally increase the amount of exercise done, inevitably participants will occasionally come shy of their monthly quota. The gym club could use these funds to offset the cost of offering the discounted plan and of any upkeep costs associated with increased use of the gym.

This approach assumes that people who visit the gym actually exercise. Although we think this is plausible—indeed, often the hardest part of exercising is overcoming inertia to get to the gym in the first

place—the scheme could be combined with a lottery incentive to assure increased exercise. Work-out machines such as treadmills and ellipticals could provide payouts after a certain number of paces. The user exercising at the time the machine "hits the jackpot" would earn a prize. To make this lotterylike incentive more enticing, rewards deemed particularly attractive to the exerciser, such as massages and other spa treatments, could be used instead of monetary payments. It is conceivable that such a program would increase patronage of a particular facility, in turn boosting revenue that would more than offset the cost of its implementation.

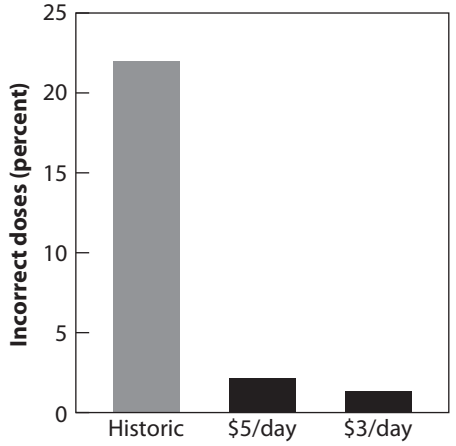
MEDICATION ADHERENCE

Poor adherence to certain prescription medications is common despite their manifest benefits. When taken properly, the drug warfarin, for example, reduces the risk of strokes by 68% overall and by 85% in patients older than 75 who have at least one other risk factor (Fuster et al., 1981; Laupacis et al., 1998; Petersen et al., 1989). Warfarin has been shown to be superior to aspirin. Because of poor compliance, however, these benefits are often not realized (Ansell et al., 1997; Cheng, 1997; Go et al., 1999; Kutner, Nixon, and Silverstone, 1991). One recent cohort study found that 40% of subjects missed 20% or more of their warfarin doses (Kimmel et al., 2007). Moreover, because warfarin taken incorrectly can be either ineffective or lead to a higher risk of bleeding and death, it has been estimated that between 45%–84% of patients with atrial fibrillation (an irregular heart beat that is a common indication for warfarin) and no contraindications for warfarin use did not receive the drug, placing them at a stroke risk several times greater than if properly anticoagulated with warfarin. This, indeed, may be the greatest negative consequence of poor adherence. The combination of proven benefits with low compliance rates in existing programs and low prescribing rates by physicians points to the need for new approaches for improving adherence. Enter behavioral economics.

In an attempt to improve adherence to warfarin regimens, Volpp and coauthors (Volpp et al., 2008) conducted pilot studies that tested the feasibility and potential effectiveness of a novel approach involving daily lottery incentives. In addition to drawing on behavioral economics, the intervention also makes use of a new technology—a computerized pillbox—that enhances the scalability of the approach.

In each of two pilots, ten patients on warfarin were provided with an Informedix Med-eMonitor System, which has a display screen and separate pill compartments that are labeled for each dose of the medication. Each device was programmed to communicate by telephone with the study's administrator. Participants were enrolled in a daily lottery; the expected value was $5 per day for the first pilot study,[4] and $3 per day for the second. Although patients were enrolled in the lottery each day that they were instructed to take a pill, they were eligible to receive any earnings only if the Med-eMonitor had conveyed that they opened the appropriate pill compartment. The Med-eMonitor also was programmed to provide a daily reminder chime as well as a message that asked whether they had taken their medication.

The primary outcome was patient adherence and was calculated as "mean correct patient pill taking" based on the percentage of days in which each patient opened the correct compartment. In the first pilot study ($5 per day expected value of lottery), 979 patient-days of warfarin use were recorded. Over this period, the mean correct pill taking was 97.7%, or only 2.3% incorrect pills, compared with a historic mean of 22% incorrect pill taking in this clinic population (fig. 21.1). Mean adherence ranged from 92% to 100% per patient. In the second pilot study ($3 per day expected value of lottery), an additional 10 patients contributed a total of 813 days of warfarin use. Mean adherence was 98.4% (only 1.6% incorrect pills taken) and ranged from 92.1% to 100%, similar to the $5-per-day pilot (fig. 21.1). Although opening pill compartments is an imperfect measure of pill taking (since patients could open the compartment but not take the pill), measurements of patients' blood coagulation rates supported the conclusion that the lottery intervention helped. In the first pilot study, the proportion of out-of-range blood coagulation rates decreased from 35.0% prepilot to 12.2% postpilot, a 65.2% improvement, and in the second pilot study, the blood coagulation rates that were out



21.1. Adherence under lotteries compared to historic controls.

of range decreased from 65.0% to 40.4%, a 37.9% improvement.

Similar to the incentive conditions of the weight-loss study, this study illustrates how a number of insights from behavioral economics—the importance of frequent feedback and incentives, the greater motivational power of lotteries of similarly valued certain payments, and the motivating force of anticipated regret—can be used to help people adhere to their medication regimens. These approaches have great potential to improve health behaviors but need to be systematically tested in a variety of clinical contexts and health delivery settings (Volpp et al., 2009).

Moving beyond the Individual: Applications at the Societal Level

While all of the applications discussed above have focused on helping individuals to help themselves, we will now explore how decision errors can be channeled instead to promote the public good. We will use the problems of global warming, charitable giving, and international conflicts to illustrate how our ideas can be applied to the public at large.

Global Warming

Although initiatives to increase individuals' consciousness about their contributions to global warming could have some impact, they are likely to, at most, make a small dent in the problem. Any serious solution to the problem is going to have to involve changes in prices, which could in some cases be enacted through taxes on energy use or subsidies for conservation. Thus, if gasoline were much more expensive in the United States, inevitably people would switch to more-fuel-efficient cars, and, in the long run, would be likely to alter their lifestyles—for example, by using public transportation or moving closer to their workplaces—in ways that reduced fuel consumption and hence emissions. However, the central point of this section is that monetary incentives of a given magnitude can have a greater or smaller impact on behavior depending on how they are implemented.

Thus, for example, insights from behavioral economics could be used to stimulate the use of public transportation. Funds currently allocated toward advertising could be redirected to a lottery-based incentive scheme. An electronic transportation pass card with a unique identifying number would be scanned each time a rider used the system. Every day, one pass-card number would be drawn, the bearer of which would receive a large prize *if she rode the system that day*. Such a policy would capitalize on the tendency to overweight small probabilities, because people would be lured into using public transportation by the small chance of winning a prize. In the same way that the weight-loss and warfarin interventions leveraged incentives by playing on regret aversion, riders in this program could be informed if their number was drawn on a day in which they did not use the system. This program could be entirely voluntary: consumers would not be obligated to participate, but we suspect that people would rather be enrolled than not.[5]

As another application of reducing transportation emissions, Greenberg (2005) discussed how mental accounting concepts can be applied in designing pay-per-mile auto insurance products. He outlined the pay-as-you-drive-and-you-save (PAYDAYS) insurance program that features individualized premiums based not on the calendar year, but on the miles a person drives. Motorists thus have the incentive of saving money on insurance by driving less. The basic premise of mental accounting is that consumers categorize their spending into separate segments or "budgets." The PAYDAYS program capitalizes on the notion that reducing the size of the insurance budget would result in decreased driving. Greenberg showed how the effect of mental accounting on reducing driving can be enhanced by applying additional behavioral economics insights, some of which we outline here and supplement with our own ideas.

FRAMING

Consumers are charged a surcharge for additional miles rather than a rebate for driving fewer than the specified number of miles. It follows from prospect theory that the former frame, in which additional driving is treated as a loss, would be more effective at reducing driving than the latter, which treats reduced driving as a windfall (gain).

PAIN OF PAYING

Research on the pain of paying (Prelec and Loewenstein, 1998) suggests that people would curtail their driving as insurance payments draw near in time because the cost of driving is salient around the time when the person makes the payment. Reduced driving may result immediately after payment while the pain of paying is still felt. This effect is bolstered by requiring consumers to make PAYDAYS payments frequently (narrow bracketing). Moreover, the pain of paying could be accentuated by equipping cars with taxi-like meters that make salient the cost of driving, much like some hybrid cars feature prominently displayed monitors indicating fuel consumption levels.

OVEROPTIMISM

The PAYDAYS program may appear especially attractive to those who overestimate their ability to cut their mileage. Such individuals may sign up for the lowest rates—rates that allow the least mileage and impose the heaviest fines on mileage overage. Then, similar to the weight-loss study discussed earlier, by virtue of having committed themselves to being virtuous and wanting to avoid fines, such individuals would be highly motivated to reduce their mileage.

In sum, by exploiting decision errors, pay per mile, which already is a good idea, could be made even more effective in achieving the goal of reducing miles driven, fuel used, and emissions.

These ideas alone would clearly be insufficient to have much of an impact on the problem of global warming. To be truly effective, they need to be paired with sweeping policy changes that increase the price of products such as gasoline to reflect their true environmental and social costs. However, these two examples illustrate an important point: that insights from behavioral economics could be applied to, in effect, supercharge such policy changes. For a more thorough discussion of how principles from psychology, behavioral economics, and behavioral decision research both contribute to and reduce global warming (and more generally, to improve environmental policy), see Weber (this volume).

Charitable Giving

In his influential book, *Living High and Letting Die* (1996), Peter Unger contrasted two scenarios a person might face. In the first, a man by the side of the road has a deep leg wound and needs immediate transportation to the hospital to avoid losing the leg. A person driving by considers helping and realizes that the blood from the victim's wound will ruin his fine leather seating and cost him $5,000. In the second scenario, the driver receives a letter from UNICEF that requests a $100 donation and informs the recipient, accurately, that unless he sends the check, several children who could be saved will instead die. The contrast is instructive because most people would harshly judge an individual who failed to help the man in the first scenario; yet the failure of so many of us to send the $100 (as detailed in the second scenario) is in fact far more egregious on a variety of dimensions. If we would condemn the driver who failed to stop in the first scenario, it follows that those of us who are "living high" yet failing to donate a large fraction of our resources to those much less fortunate than ourselves, are making a moral error.

If affluent people are not giving as much as they, in some sense, *should*, what is responsible for the shortfall of generosity? One important cause is what Thomas Schelling referred to as the identifiable victim effect: people respond more emotionally and sympathetically to identifiable individuals than to statistics. In one study of the identifiable victim effect (Small and Loewenstein, 2003), sympathy was measured by asking participants who had received $10 how much (if any) of the money they would donate to a victim, an individual who had also received $10 but had been randomly selected to lose it. Each participant drew a number from a hat, and this number represented the victim to whom they could donate. Critically, participants stated their willingness to donate either before (unidentified victim) or after (identified victim) drawing the number. Donations were about twice as high in the identified (postdraw) than the unidentified (predraw) condition. That the "identified" victim was merely a number provides an especially powerful demonstration of the effect.

Real-world paradigmatic examples of the identifiable victim effect include Jessica McClure, a girl in Texas who fell into a well and received an outpouring of sympathy and aid, and a whale that accidentally swam up the Thames River and died in close proximity to London's millions. While McClure and the whale received a tremendous amount of attention, sympathy, and aid, the millions of girls who die each year worldwide from malnutrition, malaria, and dysentery, as well as the whales that die from whaling or from the pollution of the worlds' oceans, get far less sympathy and, more important, far less help.

The identifiable victim effect is only one of a number of patterns that can be observed in charitable giving that are not consistent with standard accounts of rational choice. More generally, we know that victims who are closer in time and space or who are visible evoke greater sympathy, and that knowing the victim's story or even being exposed to the right type of music can enhance sympathy.

Is it possible to use decision errors to increase charitable giving? The answer is that it is not only possible but also a widespread practice. For example, using decision errors to boost donations is the very basis of sponsor-a-child programs. By tying donations to specific children, these programs capitalize on identifiable victim effects (see Kogut and Ritov, 2005). Given that we suspect (and hope) that fundraising tactics tying donations to specific children in fact spread the resources more widely, such tactics employ decision errors to turn a second-best situation (a small number of children get disproportionate support, while others languish) into a situation that is closer to first-best (a larger number of children get

more evenhanded support, playing on the donors' tendency to be more generous toward individuals) (see Small, Loewenstein, and Strnad, 2006, for an extended discussion of this point). In addition, potential donors are typically asked to sponsor a child for "pennies a day." Asking for small but frequent donations uses the peanuts effect and narrow framing to mitigate the donor's money loss, in turn fostering donations. This is consistent with Gourville's (1998) explanation of the successes of public radio campaigns.

There are many other ways in which charitable organizations could leverage decision errors to boost donations. For example, in the spirit of Thaler and Benartzi's (2004) Save More Tomorrow program to increase employee saving, charitable organizations could launch a "donate more tomorrow" campaign. Committing to donate more in the future is more palatable than donating now due to present-biased preferences. The common practice of having donors provide "pledges" may in fact play on such psychological mechanisms.

Anchoring and insufficient adjustment (Tversky and Kahneman, 1974) can also be used to facilitate charitable giving. Indeed, salesmen capitalize on a variant of the phenomenon the "door-in-the-face" effect, wherein a very expensive product is initially suggested to the customer. Though people usually refuse the product, they often buy something more expensive than they would have had they not been presented with the initial anchor. While asking people to donate a huge sum of money upfront runs the risk of annoying people, it also may make it more likely that they will agree to donate a smaller amount.

Shang and Croson (2006) used a type of social-comparison-based anchoring manipulation to increase over-the-phone donations to a public radio station. They found that simply mentioning that an individual contacted previously had donated a large amount increased the magnitude of the focal donor's donation. A more subtle manipulation that would be worth testing would attempt to anchor potential donors on a truly arbitrary, but high, number. For example, the American Cancer Society could ask potential donors an initial question, How much do you think it would be worth to the country to cure cancer? Based on research showing the susceptibility of valuations of anchors, even completely irrelevant ones, such a procedure is likely to boost donations.

International Disputes

When traveling through the bucolic areas in which wars seem often to be fought, one cannot help but be impressed by the contrast between the present and past. Families, towns, cities and even countries get torn apart by, as Shakespeare so aptly expressed it, the "dogs of war"—by destructive passions that sweep through populations like wildfire. Much like individuals who commit, and later pay the price for, crimes of passion, those caught up in mass hostilities often look back on their own feelings and behavior with perplexity, wondering how they could have acted as they did.

Wars, like individual self-destructive behavior, are often, prosaically, the product of individual-level irrationality. Although sometimes orchestrated or at least encouraged by those with "rational" economic interests in fomenting conflict, most wars are associated with a variety of decision errors. For example, people are often overconfident about their likelihood of prevailing, as was true at the beginning of World War I, when citizens of countries on both sides of the dispute anticipated quick victory for their own side. More important, perhaps, the passions of the moment tend to produce a variety of judgmental and motivational distortions (Loewenstein, 1996), such as a powerful motivation for immediate action (e.g., a need to act quickly rather than, for example, opt for diplomacy or the gradual effects of economic sanctions), dramatic self-serving biases when it comes to evaluating fairness, insensitivity to variations in probabilities, and extremes of sympathy, antipathy, and callousness (see Lobel and Loewenstein, 2005). Is it possible that judgmental biases can be harnessed in opposition to such effects?

As alluded to above, disputes not only tend to be the product of self-serving appraisals of the situation, but also tend to usher forth even more dramatically self-serving appraisals. In the heat of war, almost everything one's own side does is seen as benign and fair, whereas almost anything one's opponent does is interpreted in a much harsher light. As noted in the section on self-serving fairness bias, although such self-serving appraisals typically exacerbate conflict, they can also be used to enable the parties to agree on a common and respected third party to aid in the resolution of conflicts. Hence, the same self-serving bias that contributes to disputes can also be used to resolve them. For a discussion of additional ways in which engaging a third party can aid in international conflict resolution, see Ross (this volume).

Conclusion

When scrutinized superficially, the idea of using decision errors to help people might appear distasteful and misguided. Why should people have to be "tricked" into acting in their own self-interest? A more nuanced

perspective would view such uses of decision errors as a matter of balancing a playing field in which numerous corporate entities exploit decision errors in their efforts to compete in the marketplace.

There are a wide range of economic interests that exploit mistakes that consumers make (see Issacharoff and Delaney, 2006; Loewenstein and Haisley, 2008; Loewenstein and O'Donoghue, 2006). Credit card companies lure consumers with "teaser" rates that play on their naivety about their future propensity to go into debt.[6] Fast-food restaurants offer "meal deals" that would not be nearly as attractive if consumers entered calories and health consequences into their decision calculus. Cigarette and alcohol sellers broadcast ads that cast the poisoning of one's own body as a romantic, sophisticated, activity. Banks make an increasing fraction of their profits from overdraft fees that consumers do not pay much attention to when they choose where to open an account and then get "stung" by. Mortgage companies encourage consumers to assume loans they cannot afford, then the companies support legislation that makes it more difficult to declare bankruptcy and walk away from one's debts. Even states get into the game of playing on decision errors, marketing lottery tickets that return approximately 45 cents on the dollar that they sell disproportionately to those least able to afford them. The associated marketing efforts encourage simplistic assessments of probabilities—for example, the ubiquitous "you can't win if you don't play." There are many more such examples.

With the possible exception of states, which arguably should not be in the business of exploiting poor people, these economic entities are not inherently evil; they are just competing in the marketplace. If some bank or mortgage company failed to exploit consumer errors, and its competitors did, it would lose profits and risk going out of business. When consumers make systematic errors, and one can no longer assume that they are fully capable of taking actions consistent with their self-interest, there is a very real possibility that the "invisible hand" of the market will lead to the opposite result that Adam Smith envisioned.

In the best of all worlds, we could rely on the inherent rationality of individuals to help guide them through the shoals of capitalist and state enterprises that play on their biases and irrationalities. As the examples we have highlighted in this chapter suggest, however, the outcome of such a laissez-faire approach is clearly suboptimal. In the world we live in, in contrast, there are many adverse consequences of leaving consumers to fend for themselves. Harnessing the same errors that are regularly used to exploit consumers to instead help them could make many people better off.

Notes

1. Another closely related example that also involves overconfidence comes from the work of Benabou and Tirole (2002). They discuss how overconfidence in one's own abilities can in some cases counteract the reluctance, due to present-biased preferences, to engage in risky endeavors that involve an immediate outlay of effort for a delayed benefit. If people overestimate their chances of success, they may make the effort when the immediacy of costs would otherwise deter them from doing so.

2. The peanuts effect is closely related to the marketing ploy of framing costs in terms of "pennies a day" (see Gourville, 1998).

3. To deal with this problem, Gale, Gruber, and Orszag (2006) propose an alternative to the current tax-deduction-based system, which provides disproportionate benefits to savers who are in high tax brackets. In their proposal, the government would provide a 30% match to all households making a qualified contribution to a 401(k) plan or IRA account.

4. Due to a clerical error, the expected value was greater than intended. Subjects won $10 if either of their digits matched with either of the digits drawn for that day, doubling the likelihood of winning $10 above what we intended for an expected value of $5 per day. Rather than ending the trial when we discovered the error, we completed it and started a new trial with another 10 patients and the lottery implemented correctly with an expected value of $3 per day.

5. Although one might fear that car manufacturers would offer a similar program if they saw that it worked, current laws in the United States do place some restrictions on commercial entities from offering lotteries that are contingent on product usage (albeit, seemingly, mainly from offering such lottery-linked products via mail solicitations).

6. It is unlikely that these types of marketing practices are going to be regulated; if anything, there has been a tendency to move in the opposite direction—for example, with recent legislation that permits credit card accounts in which an individual's tax-free retirement savings serves as collateral.

References

Ainslie, G. (1975). Specious reward: A behavioral theory of impulsiveness and impulse control. *Psychological Bulletin*, *82*(4), 463–496.

Ansell, J. E., Buttaro, M. L., Thomas, O. V., and Knowlton, C. H. (1997). Consensus guidelines for coordinated outpatient oral anticoagulation therapy management. Anticoagulation Guidelines Task Force. *Annals of Pharmacotherapy*, *31*(5), 604–615.

Babcock, L., Loewenstein, G., Issacharoff, S., and Camerer, C. (1995). Biased judgments of fairness in bargaining. *American Economic Review*, *85*(5), 1337–1342.

Bartlett, J., Miller, L., Rice, D., and Max, W. (1994). Medical care expenditures attributable to cigarette smoking: United States. *Morbidity and Mortality Weekly Report*, *43*, 469–472.

Becker, G. S., and Murphy, K. M. (1988). A theory of rational addiction. *Journal of Political Economy*, *96*(4), 675–700.

Benabou, R., and Tirole, J. (2002). Self-confidence and personal motivation. *Quarterly Journal of Economics*, *117*(3), 871–915.

Benartzi, S., and Thaler, R. (1995). Myopic loss aversion and the equity premium puzzle. *Quarterly Journal of Economics*, *110*(1), 73–92.

Bucks, B. K., Kennickell, A. B., and Moore, K. B. (2006). Recent changes in U.S. family finances: Evidence from the 2001 and 2004 Survey of Consumer Finances. Federal Reserve Bulletin, *92*. Retrieved from http://www.federalreserve.gov/Pubs/OSS/oss2/2004/bull0206.pdf

Buehler, R., Griffin, D., and Ross, M. (1994). Exploring the planning fallacy: Why people underestimate their task completion times. *Journal of Personality and Social Psychology*, *67* (33), 366–381.

———. (2002). Inside the planning fallacy: The causes and consequences of optimistic time predictions. In T. Gilovich, D. Griffin, and D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 250–270). Cambridge: Cambridge University Press.

Bulkeley, W. M. (1998, February 10). Rebates' secret appeal to manufacturers: Few consumers actually redeem them. *Wall Street Journal*, p. B1.

Camerer, C., and Ho, T.H. (1999). Experience-weighted attraction learning in normal form games. *Econometrica*, *67*, 837–874.

Camerer, C., Issacharoff, S. Loewenstein, G., O'Donoghue, T. and Rabin, M. (2003). Regulation for conservatives: Behavioral economics and the case for "asymmetric paternalism." *University of Pennsylvania Law Review*, *1151*(3), 1211–1254.

Cheng, T. O. (1997). Underuse of warfarin in atrial fibrillation. *Archives of Internal Medicine*, *157*(13), 1505.

Chiquette, E., Amato, M. G., and Bussey, H. I. (1998). Comparison of an anticoagulation clinic with usual medical care: Anticoagulation control, patient outcomes, and health care costs. *Archives of Internal Medicine*, *158*(15), 1641–1647.

Choi, J. J., Laibson, D., and Madrian, B. C. (2005). *$100 bills on the sidewalk: suboptimal investment in 401(k) plans.* NBER Working Paper 11554. National Bureau of Economic Research.

Colcombe, S., and Kramer, A. F. (2003). Fitness effects on the cognitive function of older adults: A meta-analytic study. *Psychological Science*, *14*(2), 125–130.

Consumer Federation of America (2003, May 13). *Survey finds growing concern about personal finances, especially among the young and the least affluent.* Retrieved from http://www.consumerfed.org/pressreleases/335

Della Vigna, S., and Malmendier, U. (2006). Paying not to go to the gym. *American Economic Review*, *96*(3), 694–719.

Epley, N., and Dunning, D. (2000). Feeling "holier than thou": Are self-serving assessments produced by errors in self- or social prediction? *Journal of Personality and Social Psychology*, *79*(6), 861–875.

Flegal, K. M., Graubard, B. I., Williamson, D. F., and Gail, M. H. (2005). Excess deaths associated with underweight, overweight, and obesity. *Journal of the American Medical Association*, *293*(15), 1861–1867.

Folkins, C. H., and Sime, W. E. (1981). Physical fitness training and mental health. *American Psychologist*, *36*(4), 373–389.

Frederick, S., Loewenstein, G., and O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature*, *40*, 351–401.

Fuster ,V., Gersh, B. J., Giuliani, E. R., Tajik, A. J., Brandenburg, R. O., and Frye, R. L. (1981). The natural history of idiopathic dilated cardiomyopathy. *American Journal of Cardiology*, *47*(3), 525–531.

Gale, W. G., Gruber, J., and Orszag, P. R. (2006). *Improving opportunities and incentives for saving by middle-and low-income households.* Hamilton Project discussion paper. April. Washington.

Genesove, D., and Mayer, C. (2001). Loss aversion and seller behavior: Evidence from the housing market. *Quarterly Journal of Economics*, *116*, 1233–1260.

Glaeser, E. (2006). Paternalism and psychology. *University of Chicago Law Review*, *73*(1), 133–156.

Gneezy, U., and Potters, J. (1997). An experiment on risk taking and evaluation periods. *Quarterly Journal of Economics*, *112*(2), 631–645.

Go, A. S, Hylek, E. M., Borowsky, L. H., Phillips, K. A., Selby, J. V., and Singer, D. E. (1999). Warfarin use among ambulatory patients with nonvalvular atrial fibrillation: The anticoagulation and risk factors in atrial fibrillation (ATRIA) study. *Annals of Internal Medicine*, *131*(12), 927–934.

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist*, *54*(7), 493–503.

Gourville, J. T. (1998). Pennies-a-day: The effect of temporal reframing on transaction evaluation. *Journal of Consumer Research*, *24*(4), 395–403.

Greenberg, A. (2005). Applying mental accounting concepts in designing pay-per-mile auto insurance products. Federal Highway Administration, Office of Policy. Washington, DC. Retrieved from http://www.trb-pricing.org/docs/06-2967.pdf

Guillen, M., and Tschoegl, A. (2002). Banking on gambling: Banks and lottery-linked deposit accounts. *Journal of Financial Services Research*, *21*, 219–231.

Haisley, E., Mostafa, R., and Loewenstein, G. (2008). Subjective relative income and lottery ticket purchases. *Journal of Behavioral Decision Making*, *21*, 283–195.

Halpern, S. D., Ubel, P. A., and Asch, D. A. (2007). Harnessing the power of default options to improve health care. *New England Journal of Medicine*, *357*(13), 1340–1344.

Hamermesh, D. S., and Soss, N. M. (1974). An economic theory of suicide. *Journal of Political Economy*, *82*(1), 83–98.

Herrnstein, R. J., and Prelec, D. (1992). Melioration. In G. Loewenstein and J. Elster (Eds.), *Choice over time* (pp. 235–263). New York: Russell Sage Foundation.

Hughes, J. R. (2003). Motivating and helping smokers to stop smoking. *Journal of General Internal Medicine*, *18*(12), 1053–1057.

Issacharoff, S., and Delaney. E. F. (2006). Credit card accountability. *University of Chicago Law Review*, *73*, 157–182.

Jackevicius, C. A., Mamdani, M., and Tu, J. V. (2002). Adherence with statin therapy in elderly patients with and without acute coronary syndromes. *Journal of the American Medical Association*, *288*(4), 462–467.

Johnson, E. J., and Goldstein, D. (2003). Do defaults save lives? *Science*, *302*(5649), 1338–1339.

Kahneman, D., Knetsch, J. L., and Thaler, R. H. (1991). The endowment effect, loss aversion, and status quo bias: Anomalies. *Journal of Economic Perspectives*, *5*(1), 193–206.

Kahneman, D., and Lovallo, D. (1993). Timid choices and bold forecasts: A cognitive perspective on risk taking. *Management Science*, *39*(1), 17–31.

Kahneman, D., and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–291.

Kimmel, S. E., Chen, Z., Price, M., Parker, C. S., Metlay, J. P., Christie, J. D., et al. (2007). The influence of patient adherence on anticoagulation control with warfarin: Results from the International Normalized Ratio Adherence and Genetics (IN-RANGE) Study. *Archives of Internal Medicine*, *167*(3), 229–235.

Kivetz, R., Urminsky, O., and Zheng, Y. (2006). The goal-gradient hypothesis resurrected: Purchase acceleration, illusionary goal progress, and customer retention. *Journal of Marketing Research*, *43*(1), 39–58.

Kogut, T., and Ritov, I. (2005). The "identified victim" effect: An identified group, or just a single individual? *Journal of Behavioral Decision Making*, *18*, 157–167.

Kutner, M., Nixon, G., and Silverstone, F. (1991). Physicians' attitudes toward oral anticoagulants and antiplatelet agents for stroke prevention in elderly patients with atrial fibrillation. *Archives of Internal Medicine*, *151*(10), 1950–1953.

Lambrecht, A., and Skiera, B. (2006). Paying too much and being happy about it: Existence, causes and consequences of tariff-choice biases. *Journal of Marketing Research*, *43*, 212–223.

Laupacis, A., Albers, G., Dalen, J., Dunn, M. I., Jacobson, A. K., and Singer, D. E. (1998). Antithrombotic therapy in atrial fibrillation. *Chest*, *114*(5), 579S–589S.

Lipsey, R. G., and Lancaster, K. (1956). The general theory of second best. *Review of Economic Studies*, *24*(1), 11–32.

Lobel, J., and Loewenstein, G. (2005). Emote control: The substitution of symbol for substance in foreign policy and international law. *Chicago Kent Law Review*, *80*(3), 1045–1090.

Loewenstein, G. (1992). The fall and rise of psychological explanation in the economics of intertemporal choice. In G. Loewenstein and J. Elster (Eds.), *Choice over time* (pp. 3–34). New York: Russell Sage.

———. (1996). Out of control: visceral influences on behavior. *Organizational Behavior and Human Decision Processes*, *65*, 272–292.

Loewenstein, G., and Angner, E. (2003) Predicting and indulging changing preferences. In G. Loewenstein, D. Read, and R. Baumeister (Eds.), *Time and decision: Economic and psychological perspectives on intertemporal choice* (pp. 351–391). New York: Russell Sage.

Loewenstein G., Brennan, T., and Volpp, K. G. (2007). Asymmetric paternalism to improve health behaviors. *Journal of the American Medical Association*, *298*(20), 2415–2417.

Loewenstein, G., and Haisley, E. (2008). The economist as therapist: Methodological issues raised by "light" paternalism. In A. Caplin and A. Schotter (Eds.), *The foundations of positive and normative economics: A handbook* (pp. 210–245). Oxford: Oxford University Press.

Loewenstein, G., and O'Donoghue, T. (2006). We can do this the easy way or the hard way: Negative emotions, self-regulation, and the law. *University of Chicago Law Review*, *73*, 183–206.

Loewenstein, G., O'Donoghue, T., and Rabin, M. (2003). Projection bias in predicting future utility. *Quarterly Journal of Economics*, *118*, 1209–1248.

Loibl, C., Haisley, E., and Loewenstein, G. (2011). *Testing strategies to increase saving and retention in individual development account programs.* Manuscript in preparation.

Madrian, B. C., and Shea, D. F. (2000). The power of suggestion: Inertia in 401(k) participation and savings behavior. NBER Working Paper Series 7682. National Bureau of Economic Research.

Markowitz, H. (1952). The utility of wealth. *Journal of Political Economy*, *60*, 151–158.

Mokdad, A. H., Marks, J. S., Stroup, D. F., and Gerberding, J. L (2000). Actual causes of death in the United States. *Journal of the American Medical Association, 291*(10), 1238–1245.

Murphy, K. (2006, October 26). Keynote management presentation. *Forging a societal action plant in preventing childhood obesity around the world*. McGill Integrative Health Challenge Think Tank. Montreal.

Nordgren, L. F., van der Pligt, J., and van Harreveld, F. (2008). The instability of health cognitions: Visceral states influence self-efficacy and related health beliefs. *Health Psychology*, *27*(6), 722–727.

Odean, T. (1998). Are investors reluctant to realize their losses? *Journal of Finance*, *53*(5), 1775–1798.

O'Donoghue, T., and Rabin, M. (1999). Doing it now or later. *American Economic Review*, *89*(1), 103–124.

———. (2000). The economics of immediate gratification. *Journal of Behavioral Decision Making*, *13*(2), 233–250.

Petersen, P., Boysen, G., Godtfredsen, J., Andersen, E. D., and Andersen, B. (1989). Placebo-controlled, randomised trial of warfarin and aspirin for prevention of thromboembolic complications in chronic atrial fibrillation. The Copenhagen AFASAK study. *Lancet*, *1*(8631), 175–179.

Polivy, J., and Herman, P. (1992). Undieting: A program to help people stop dieting. *International Journal of Eating Disorders*, *11*(3), 261–268.

Prelec, D., and Loewenstein, G. (1991). Decision making over time and under uncertainty: A common approach. *Management Science*, *37*, 770–786.

———. (1998). The red and the black: Mental accounting of savings and debt. *Marketing Science*, *17*, 4–28.

Read, D., Loewenstein, G., and Rabin, M. (1999) Choice bracketing. *Journal of Risk and Uncertainty*, *19*, 171–197.

Read, D., and van Leeuwen, B. (1998). Predicting hunger: The effects of appetite and delay on choice. *Organizational Behavior and Human Decision Processes*, *76*, 189–205.

Ross, L., and Ward, A. (1996). Naive realism in everyday life: Implications for social conflict and misunderstanding. In T. Brown, E. S. Reed, and E. Turiel (Eds.), *Values and knowledge. The Jean Piaget Symposium Series* (pp. 103–135). Hillsdale, NJ: Erlbaum.

Sabini, J., and Silver, M. (1982). *Moralities of everyday life*. Oxford: Oxford University Press.

Samuelson, W., and Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, *1*, 7–59.

Satcher, D. (2006). The prevention challenge and opportunity. *Health Affairs*, *25*, 1009–1011.

Schroeder, S. A. (2007). Shattuck Lecture. We can do better–improving the health of the American people. *New England Journal of Medicine*, *357*(12), 1221–1228.

Shang, J., and Croson, R. (2006). The impact of social comparisons on nonprofit fundraising. *Research in Experimental Economics*, *11*, 143–156.

Shefrin, H., and Statman, M. (1985). The disposition to sell winners too early and ride losers too long. *Journal of Finance*, *40*, 777–790.

Slovic, P. (2001). Cigarette smokers: Rational actors or rational fools? In P. Slovic (Ed.), *Smoking: Risk, perception, and policy* (pp. 97–126). Thousand Oaks, CA: Sage.

Small, D. A., and Loewenstein, G. (2003). Helping *a* victim or helping *the* victim: Altruism and identifiability. *Journal of Risk and Uncertainty*, *26*, 5–16.

Small, D., Loewenstein, G., and Strnad, J. (2006). Statistical, identifiable and iconic victims and perpetrators. In E. McCaffery and J. Slemrod (Eds.), *Behavioral public finance: Toward a new agenda*. New York: Russell Sage Foundation Press.

Thaler, R. H. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior and Organization*, *1*, 39–60.

———. (1985). Mental accounting and consumer choice. *Marketing Science*, *4*, 199–214.

Thaler, R. H., and Benartzi, S. (2004). Save more tomorrow: Using behavioral economics to increase employee saving. *Journal of Political Economy*, *112*(1), S164–S187.

Thaler, R. H., and Sunstein, C. R. (2003). Libertarian paternalism. *American Economic Review*, *93*(2), 175–179.

Thaler, R. H., Tversky, A., Kahneman, D., and Schwartz, A. (1997). The effect of myopia and loss aversion on risk taking: An experimental test. *Quarterly Journal of Economics*, *112*, 647–661.

Tufano, P. (2008). Saving whilst gambling: An empirical analysis of UK premium bonds. *American Economic Review*, *98*(2), 321–26.

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

———. (1991). Loss aversion in riskless choice: A reference-dependent model. *Quarterly Journal of Economics*, *106*(4), 1039–1061.

Unger, P. K. (1996). *Living high and letting die: Our illusion of innocence*. New York: Oxford University Press.

U.S. Census Bureau. (2003). Net worth and asset ownership of households: 1998 and 2000. Publication P70-88. Retrieved from http://www.census.gov/prod/2003pubs/p70-88.pdf

Viscusi, W. K. (1992). *Smoking: Making the risky decision*. New York: Oxford University Press.

Volpp, K. G., John, L. K., Troxel, A. B., Norton, L., Fassbender, J., and Loewenstein, G. (2008). Financial incentive-based approaches for weight loss: A

randomized trial. *Journal of the American Medical Association*, *300*(22), 2631–2637.

Volpp, K. G., Loewenstein, G., Troxel, A. B., Doshi, J., Price, M., Laskin, M., and Kimmel, S. E. (2008). A test of financial incentives to improve warfarin adherence. *Biomedical Central: Health Services Research*, *8*, 272–278.

Volpp, K. G., Pauly, M. V., Loewenstein, G., and Bangsberg, D. (2009). P4P4P: An agenda for research on pay for performance for patients. *Health Affairs*, *28* (1), 206–214.

Weber, B. J., and Chapman, G. B. (2005). Playing for peanuts: Why is risk seeking more common for low-stakes gambles? *Organizational Behavior and Human Decision Processes*, *97*, 31–46.

Weber, M., and Camerer, C. (1998). The disposition effect in securities trading: An experimental analysis. *Journal of Economic Behavior and Organization*, *33*, 167–184.

Woolf, S. H. (2007). Potential health and economic consequences of misplaced priorities. *Journal of the American Medical Association*, *297*(5), 523–525.

CHAPTER 22

Doing the Right Thing Willingly

Using the Insights of Behavioral Decision Research for Better Environmental Decisions

ELKE U. WEBER

Policy makers from local to supranational levels are being asked to address behavior that impacts economic and social outcomes on multiple scales and, increasingly, also environmental outcomes. Attempts to reduce a country's dependence on foreign oil, for example, may generate multiple options that all satisfy this policy goal but can have varying impacts on the economic viability as well as on air quality and carbon dioxide emissions. Custodians of the water available in a system of reservoirs need to regulate release times and levels in a way that satisfies stakeholders with different needs and sources of power, while safeguarding future availability of water, given projected future upstream rainfalls. At other times, environmental issues pose themselves as the primary problem for policy intervention. For example, national legislation or supranational agreements may mandate that regional emissions of harmful substances such as sulfur or carbon dioxide be capped at specific levels, and regional or industry-specific policy makers need to generate interventions that will reduce emission-generating activities or introduce technologies that reduce emission levels of continuing activities.

These examples illustrate several important points: (1) Environmental policy decisions typically have impacts on a range of dimensions, from economic to social and environmental ones and involve trade-offs between these dimensions. (2) Many of these decisions have distributional implications and involve considerations of fairness or equity. (3) Many of these decisions involve considerable uncertainty about the likely consequences of different actions and require intertemporal trade-offs on both the cost and benefit side. (4) Implementation of such policies typically involves persuasion, for example, convincing people or groups to reduce consumption in situations where economic models of rational behavior argue against such reductions. Environmental goods such as clean air, drinkable water, species diversity, and a life-sustaining climate are common-pool resources, and rational economic analysis prescribes short-sighted, selfish depletion of such resources (or failure to invest in their upkeep to ensure their continued existence or quality) as the dominant behavior, even though more long-sighted and cooperative behavior would be socially desirable (Bowles, 2004). While most policy decisions possess these four characteristics to some extent, they seem to loom particularly large for policy decisions in the environmental domain.

Theoretical Background and Assumptions

In describing environmental decision-making processes in their possible variations, this chapter draws on theory in social cognition. Behavior is assumed to be determined by unconscious and conscious inference and decision processes, which are elicited by conditions in the external environment in combination with internal factors that include prior experience, expectations, and goals (Weber and Johnson, 2009). This body of theory is informed by insights from behavioral decision research that has documented people's limitations in attention, memory, and information processing. It is a perspective often referred to as bounded rationality (Simon, 1982). When preferences are constructed while decisions are made, the processes used to do so are different and often simpler than the as-if calculations implicitly assumed by rational-economics models of choice (Kahneman, 2003; Lichtenstein and Slovic, 2006).

The first part of the chapter will review some behavioral phenomena likely to be at play in environmentally relevant decisions that should *increase* our concern about the challenges faced by environmental policy makers beyond those already established by a rational-expectations analysis of common-pool resource dilemmas (Hardin, 1968). In particular, we will examine the negative impact of the following phenomena: (1) People lack appropriate visceral

reactions to important classes of environmental risks. (2) Cognitive and affective myopia, coupled with loss aversion, makes the immediate costs and sacrifices typically required for environmentally responsible behavior loom large, whereas future benefits have little appeal because people apply extremely high discount rates. (3) The uncertainty of future risks or benefits complicates the task even more, with ambiguity aversion and the underweighting of small probabilities in decisions based on personal experience of consequences playing important roles in people's environmentally relevant decisions. As a result, decision makers who approach environmentally relevant decisions in either an analytic or an affect-based mode will not likely voluntarily modify existing problematic behavior, for example, reduce their energy consumption.

Fortunately, this is not where the story ends. The second part of this chapter will ask whether Hardin's (1968) "tragedy" of the commons could perhaps be downgraded to a "drama" (Ostrom et al., 2002). We will see that people might be induced to act in more collective ways that also increase their own long-term individual benefits if three sources of cognitive abundance with which they are equipped are used to shape the decision environment in ways that will facilitate more environmentally sustainable behavior: (1) the multiple ways in which they can look at information (e.g., framing, mental accounting), (2) the broad range of goals (e.g., individual vs. social goals, promotion vs. prevention goals) they have that can be selectively activated, and (3) their ability to decide upon a course of action in multiple and qualitatively different ways (e.g., using habits, rules, roles, affect, and calculations).

Insights derived from these sources of cognitive abundance can guide the design of environmental policies. This might include interventions to induce the American public to implement a long list of existing energy-efficiency innovations (e.g., home insulation or different lighting technology like LED or CFL bulbs) that would result in no reduction of their standard of living, produce a net cost savings over a multiyear period, and sizably reduce U.S. energy-use and carbon dioxide emissions (Granade et al., 2009). The ways in which people process information about uncertain events removed in space and time will be discussed in the next section and may help explain why these alternative actions, which appear to be economically and environmentally dominating ("low-hanging fruit"), are not being adopted by the overwhelming majority of the American public.

Behaviors of Concern

The different ways in which people process information when making judgments or arriving at decisions have been classified into two contrasting categories, sometimes referred to as two "systems" (Chaiken and Trope, 1999; Kahneman, 2003; Sloman, 1996). The first category of processes works on the basis of temporal and spatial associations and similarity. It uses real-world experience as input. Its basic mechanisms are automatic, that is, associations are established, stored, and retrieved essentially without effort and conscious awareness. Such associative processes teach us, for example, to dislike food eaten just prior to symptoms of food poisoning and to avoid foods of similar taste or smell in the future. Associative processes map uncertain and adverse aspects of the environment into affective responses (e.g., fear, dread, anxiety) and thus represent *risk* as a *feeling* (Loewenstein et al., 2001).

Many contemporary environmental or technological risks (e.g., climate change or nuclear power) do not (yet) provide direct experience of adverse consequences most of the time, either because of successful risk management or because the adverse consequences have a small probability and often lie in the future. Such risks, based on model-based predictions, are typically communicated to the public in an abstract and symbolic way, for example, as probability distributions of possible consequences. Such information needs to be processed by the second category of processes people have at their disposal. This second class of processes works on the analytic algorithms and rules specified by normative models of judgment and choice (e.g., the probability calculus, Bayesian updating, formal logic, and utility maximization) and also on simpler versions of such algorithms that explicitly combine information. They are slower than automatic associative processes and require conscious awareness and control. The algorithms that these analytic processes implement need to be taught explicitly, and the appropriateness of their use for a given situation needs to be apparent, that is, they do not get triggered automatically.

Hardin and Banaji (this volume) similarly distinguish between visible conscious and invisible unconscious (implicit) processes. Such dual-process accounts have been very useful as a conceptual framework, although one has to be careful not to take the dichotomy too literally. While elements of the two processing systems can operate in parallel, it is unclear whether they can operate in isolation, and they also interact with each other in complex ways (Evans, 2007; Weber and Johnson, 2009). Analytic reasoning is often guided and assisted by automatic processes that include associations and affect (Damasio, 1994), and few decisions are made in a completely reflexive way. When both types of processing are in operation but their outputs disagree, the output of the associative system typically prevails, because its output has greater vividness and emotional salience.

Even in seemingly very analytic contexts, such as financial investment decisions, subjective and largely affective factors have been shown to influence perceptions of risk (Holtgrave and Weber, 1993) and the choice of investment options (Weber, Siebenmorgen, and Weber, 2005). Hersch and Viscusi (2006) connect affective factors to seemingly analytic considerations in the environmental domain, showing that national differences in worry about global warming correlate with willingness to pay more for gasoline, if such price increases would result in less harm to the environment.

Insufficient Visceral Reactions to Environmental Risks

As suggested by Peters and Slovic (2000), affect—and, in particular, negative affect—is the wellspring of action. The feeling of fear powerfully motivates us to remove ourselves from a dangerous situation (Loewenstein et al., 2001). The absence of any affective or visceral response to such environmental risks as radon contamination, coastal plains flooding, or climate change may well be responsible for the arguably less-than-optimal allocation of personal and collective resources to deal with such issues (Dunlap and Saad, 2001). Behavioral decision research over the past thirty years provides some answers as to why the general population and their public officials may show less concern about some risks considered significant by domain experts, but then overreact to other risks, which experts consider insignificant.

People's affective reactions to risky situations often do not agree with more objective measures of risk that quantify either the statistical unpredictability of outcomes or the magnitude or likelihood of adverse consequences (Sunstein, 2006 and this volume). Instead, visceral judgments of risk are determined by other (psychological) risk characteristics that elicit affective reactions as part of our evolutionary heritage. The psychological risk dimensions that strongly influence judgments of the riskiness of material risks in ways that go beyond their objective consequences (Fischhoff et al., 1978) are described by two factors (Slovic, 1997). *Dread risk*, the first factor, is experienced in the face of hazards associated with a perceived lack of control over risk exposure and with consequences that are potentially catastrophic: terrorist attacks, nuclear reactor accidents, or nerve-gas attacks. *Unknown risk*, the second factor, is associated with how much is known about the hazard, how easily exposure and adverse consequences are detectable, and whether it is natural or man-made. At the high (top) end, we find chemical hazards and radiation, which might kill exposed parties without their awareness, and DNA technology, which might have serious consequences not yet tested by time. Slovic, Lichtenstein, and Fischhoff (1984) suggest that these more affective reactions to risk are forward-looking in ways not always captured by the expected value calculations of experts based on actuarial figures or scientific models. A large accident portends possibly even larger future trouble, and concern about catastrophic potential or lack of control may play a useful societal function.

The risks analyzed to infer these psychological (more visceral than analytic) reactions were mostly technological and household health risks. It is instructive to try to place some important environmental risks into this two-dimensional space. If people conceive of climate change, for example, as a simple and gradual change on variables such as average temperatures and precipitation or the frequency or intensity of specific extreme weather events (frosts, hurricanes, or tornadoes), then the risks posed by climate change would appear to be well known and exposure, at least in principle, to be controllable at the individual level ("move from Miami to Vancouver when things get too hot or dangerous in Florida"). While some of the perceived control may be illusory, the perceived ability or inability to take corrective action is an important component of vulnerability.

The main conclusion from this section is that, without sufficiently strong visceral reactions to many environmental risks (if they are considered "natural" and well known), people may not be motivated to take corrective or evasive actions. In the section on potentially useful behavioral insights, I will argue that risks can be reframed, and environmental risks can be presented as more uncontrollable, or man-made, to activate the feelings that something is amiss, which is known to result in greater risk management.

Appeals to fear are problematic for reasons beyond the fact that people do not naturally worry about environmental risks like climate change, one such reason being that people appear to have a *finite pool of worry* (Weber, 2006). As concern about one type of risk increases, worry about other risks frequently decreases, as if people had a limited budget to spend on worry. A Pew Research Center opinion poll (2009) found that levels of concern about climate change had declined in October 2009 relative to a high in 2006 that had been maintained as late as May 2008. Presumably that decline in concern with the climate was the result of increased concern about the national and world economy and unemployment. Hansen, Marx, and Weber (2004) found evidence that was consistent with a finite pool of worry among farmers in the Argentine Pampas. As concern about climate risk increased in the course of a two-day farm decision workshop that provided information about the potential impacts of increased climate variability, concern about political risk went down (post- vs. pre-workshop) even though the level of political risk had not changed over those

two days. In addition, those who stated greater worry about political risk (either pre- or post-workshop) worried less about climate risk. If people's capacity for worry or concern is finite, then efforts to raise greater concern to motivate protective or mitigation action against some risk by, for example, providing concrete images of possible damages, come at the cost of potentially reducing concern about other risks. The finite pool of worry concept is related to, though certainly not identical to, the concept of risk homeostasis (Wilde, 1998).

Appeals to fear are also problematic because of the *single-action bias* (Weber, 1997), which is the propensity to take only a single action in response to a fear signal, even in situations where a broader set of remedies might be called for. Taking the one action to respond to a problem at hand seems to reduce or remove the feeling of worry or concern. Without the latter affective marker, motivation for further action is reduced. Weber (1997) found that Midwestern farmers engaged in only one of three plausible classes of protective actions against climate change. Hansen, Marx and Weber (2004) similarly found that farmers in Argentina employed only one of several protections against climate variability and climate change. If they had the capacity to store grain, for example, they were less likely to also irrigate and invest in crop insurance. Thus fear appeals may also backfire because they motivate people to take simpler actions than are warranted by the complexity of contemporary problems.

Cognitive Myopia, Loss Aversion, and Hyperbolic Time Discounting

Sunstein (this volume) depicts cost-benefit analysis as a solution against "misfearing," that is, against people's incorrectly calibrated reactions as described in the previous section, as well as others. "The problem of misfearing," according to Sunstein, "results from use of the availability heuristic, from informational and reputational cascades, from intense emotional reactions, from processes of reasoning in which benefits are salient but costs are not, or from miscalculating the systemic effects of one-shot interventions." However, the behavioral evidence to be presented in this section suggests that environmental decisions are problematic not just when addressed affectively but also when based on calculations that trade-off costs against benefits, outcomes against probabilities, or generally evaluate the consequences of choice options in a more analytic fashion. Sunstein's remedy may make some sense when applied to the cost-benefit analyses done by domain experts, but not to the on-the-fly (and hence more fallible) calculation-based decisions described in this section, although some of the issues (e.g., about the correct discount rate to value future costs and benefits) encountered with intuitive calculations also surface in debates about expert-based cost-benefit analyses (e.g., Weitzman, 2007)

What the behavioral regularities described in this section have in common is that they bias the analytic evaluation of choice options in environmentally impactful situations against socially responsible and long-term, individually and socially beneficial behavior, which typically involves immediate costs and sacrifices that loom large, while their much delayed and uncertain future benefits get unreasonably discounted.

COGNITIVE MYOPIA

Myopia, or shortsightedness, has been cited as an explanatory construct in the context of loss aversion, most prominently by Benartzi and Thaler (1995) in their explanation of the equity premium puzzle, that is, of the puzzling fact that investors hold bonds to the degree that they do, given that the returns on stocks are significantly larger, albeit riskier. That behavior, which is inconsistent with reasonable assumptions about risk aversion, can be explained by the assumption that investors do not apply sufficiently long time horizons to their investment decisions but, instead, compare and contrast the outcomes of risk-free and risky investment opportunities on a quarterly basis and get disproportionately agitated by losses. Such shortsightedness in their time horizon also contributes to people's reluctance to save adequately for their retirement, unless such saving is legally mandated or encouraged by nudges that take advantage of people's myopia in some form of psychological judo (Thaler et al., this volume). Failures to integrate the outcomes of a series of decisions that should be considered in combination (e.g., the returns on an investment across a series of months or the returns across all investments in one's portfolio) are another example of myopia, which focuses attention on just the most recent return or the single investment (Read, Loewenstein, and Rabin, 1999; Thaler and Johnson, 1990).

Cognitive myopia thus prevents people from accurately perceiving the future benefits of immediate costs or of reductions in immediate benefits. As a result, people fail to buy more-energy-efficient appliances or make a host of other energy efficiency investments, where higher up-front purchase costs are more than compensated for by future energy savings (Gillingham, Newell, and Palmer, 2009).

LOSS AVERSION

Loss aversion is the label given to an important property that distinguishes prospect theory (Tversky and Kahneman, 1992) from expected utility theory (von Neumann and Morgenstern, 1944), namely a much

greater (dis)utility for outcomes that are encoded as losses relative to a reference point than for outcomes of the same magnitude but encoded as gains relative to a different reference point. Loss aversion explains a broad range of choices observed in both the laboratory and the real word that deviate from the predictions of rational-economic choice theory (Camerer, 2000). Employees may be willing to forgo projected future increases in salary (forgone gains) but will fight tooth and nail to avoid any cuts in their current salary (losses). With the status quo as a very salient reference point, loss aversion makes it hard for policy makers to convince people to reduce consumption or, more generally, their standard of living below current levels. While naturally used reference points in combination with loss aversion can be problematic in a range of policy domains (see Thaler, this volume), prospect theory also provides policy makers with a design tool, namely the ability to change decision makers' reference points, with implications for the way in which outcomes get evaluated. The purchase of an insurance policy against drought by a farmer, for example, involves a sure out-of-pocket loss of money (the insurance premium) for the unsure and low-probability benefit of avoiding a much larger loss in the case of drought. Prospect theory predicts risk seeking in the domain of losses, which would mean choosing the probabilistic loss over the sure loss. Skillful insurance salespeople have long known that they need to move a farmer's reference point away from its usual position at the status quo, down to the level of the possible large loss that could be incurred in case of drought. By focusing the insuree's attention on the severity of the possible loss and the resulting consequences, all the smaller losses (including the insurance premium) are to the right of (thus less negative than) this new reference point, making this a decision in the domain of (forgone) gains, where people are known to be risk averse and will choose the sure option of buying the insurance.

Attribute framing can have similar effects. Levin and Gaeth (1988) showed that people rated the taste of minced beef higher when it was described to them as 75% lean than as 25% fat, presumably because the discrepancy between 25% and 0% fat (a relative loss) is considered more severe than the discrepancy between 75% and 100% lean (a foregone gain). A recent study showed that Republicans were much more likely to purchase a more expensive plane ticket that included a fee to compensate the carbon dioxide emissions generated by the flight when that fee was called an offset (which was presumably encoded as a foregone gain) rather than a tax (which most people, and especially Republicans, encoded as an out-of-pocket loss) (Hardisty, Johnson, and Weber, 2010).

HYPERBOLIC TIME DISCOUNTING

Future financial costs and benefits ought to be discounted in value (e.g., by the current rate of interest offered by banks), ideally by a constant amount per period of time delay, as described by an exponential discount function. Empirical research shows, however, that people apply sharp discounts to costs or benefits that will occur at some point in the future relative to obtaining them immediately (e.g., a year into the future vs. now) but discount much less when both time points are in the future, with one occurring later than the other (e.g., two years versus only one year into the future) (Loewenstein and Elster, 1992). Such behavior has been described by a hyperbolic discount function that shows its steepest decrement in current value as we defer immediate consumption (Ainslie, 1975). Actions to mitigate negative environmental consequences are unattractive within this framework because they require immediate sacrifices in consumption that are compensated only by heavily discounted and highly uncertain benefits at a much later point in time.

In many situations, including those of intertemporal choice, people do not have firmly established preferences for choice options but, instead, construct them as they go by recruiting arguments for different choice options, by examining external evidence, and by recruiting internal evidence from memory (Lichtenstein and Slovic, 2006; Payne, Bettman, and Johnson, 1993; Weber and Johnson, 2009). Trope and Liberman (2003) showed that when people recruit evidence internally, events in the future elicit different arguments for and against them than imminent events. Events in the distant future (e.g., an invitation to give a paper at a conference next summer, or the prospect of coastal flooding thirty years from now, to use an environmental example) are construed in abstract terms, whereas events close to us in time (the upcoming trip on Monday to attend the long-scheduled conference, or the prospect of a major hurricane passing through town this afternoon) are construed in very concrete terms. Abstract representations of consequences in the distant future lack the concrete associations that are connected to emotional reactions. In contrast, concrete representations of consequences in the present tend to be saturated with affective associations. This difference in the affective richness and concreteness of the representation of temporally close versus distant consequences may well lie at the root of observed problems of self-control, be they impatience and impulsivity in obtaining desirable outcomes (Laibson 1997; Mischel, Grusec, and Masters, 1969) or procrastination with undesirable tasks (O'Donoghue and Rabin, 1999). Mitigating

actions against environmental problems are often perceived as requiring the sacrifice of concrete, immediate benefits for the sake of abstract, distant goals. As will be discussed in the section on useful behavioral insights, there are other and more positive ways of framing such choices. However, when pro-environmental behaviors are framed as involving sacrifices, the strong negative affect associated with the concrete, immediate costs, the absence of feelings of worry about abstract and distant negative consequences of failures to act, and the discounting of future benefits will result in ecologically damaging consumption decisions and actions.

The preferences-as-memory framework of Weber and Johnson (2006) has examined the attentional processes and memory-retrieval operations that underlie preference construction. Under this framework, query theory (Johnson, Häubl, and Keinan, 2007) assumes that when asked to delay consumption, people first assess the evidence arguing for immediate consumption and only then assess evidence that argues for delaying consumption. Query theory postulates that, in order to help people reach a decision, evidence generated in favor of an action (e.g., immediate consumption) tends to interfere with the subsequent generation of evidence arguing against that action and for other actions. Weber et al. (2007) provided empirical support for both conjectures and succeeded in drastically reducing the intertemporal discounting in people's choice by prompting them to first generate evidence in favor of deferring consumption, followed by a prompt to generate evidence in favor of immediate consumption. Query theory thus provides policy makers with a tool that may help with the successful implementation of environmental policies as further discussed in the section "Useful Behavioral Insights," below.

Risk and Ambiguity Aversion and Small Probabilities

In addition to behavioral phenomena that influence the valuation of outcomes of different choices, there also are behavioral regularities that can bias people's evaluation of the probabilities of environmentally relevant choice options.

RISK AND AMBIGUITY AVERSION

Expected utility theory (von Neumann and Morgenstern, 1944) has been central in the analysis of choice under risk and uncertainty not only for its compelling axiomatic foundation and mathematical tractability, but also for its ability to describe a large number of economic choices (Woodward, 1998). It describes deviations from expected value maximization by postulating a nonlinear and mostly concave utility function that goes back to Bernoulli (1738/1954). Classical demonstrations referred to as the Allais (1953) and Ellsberg (1961) paradoxes have given rise to additional theoretical elaborations (Camerer 2000; McFadden 1999). The Allais paradox demonstrates the certainty effect, an important feature of prospect theory (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992). The certainty effect, namely, that certain outcomes get more decision weight than they deserve based on their likelihood of occurrence, is captured by prospect theory's probability weighting function, which has a discontinuity before the endpoints, making events that occur or do not occur for sure far more impactful than those that occur with probability .999 or .001, respectively. Because sure outcomes in environmentally relevant decisions (such as deciding between a more energy-efficient refrigerator with a higher purchase price and a cheaper refrigerator with higher energy consumption and thus more carbon dioxide emissions) tend to be on the negative or cost side (i.e., the greater purchase price), while gains (i.e., the energy savings) are delayed in time and somewhat uncertain, it is easy to see that the certainty effect may introduce yet another bias toward environmentally less responsible choices in such decisions.

The Ellsberg paradox established that decision makers distinguish between well-specified probabilities (risk) and ill-defined probabilities (uncertainty), even if the best estimates of the latter have the same value as the former. Ellsberg (1961) referred to the dislike that decision makers have for options with ill-defined probabilities as *ambiguity aversion*, and Hsu et al. (2005) recently provided neuroimaging evidence that risky and uncertain choices are processed in different brain regions. Heath and Tversky (1991) demonstrated that ambiguity aversion is not universal and, in particular, is not found in situations in which decision makers believe they have expertise in the domain of choice, preferring, for example, sports gambles with ill-defined probabilities of winning or losing to money lotteries with well-specified probabilities. Whereas one can expect to find many members of the general public who think of themselves as experts in such domains as sports or the stock market, and thus do not shy away from choice options with ill-defined probabilities, the number of people who would believe themselves to be experts in environmentally relevant technical domains (e.g., the pros and cons of hybrid electric vs. conventional gasoline engines in cars) has to be much smaller at this time. This suggests that for such decisions the uncertainty and ambiguity of positive benefits of environmentally responsible choice options will more likely be seen as a

handicap rather than an opportunity. It also suggests a perhaps unexpected advantage of educating the public about technological innovations. Such education not only provides more accurate input for people's analytic processing of environmentally relevant choice options, but will also help to remove negative affective reactions to uncertainty that are associated with not-well-understood probabilistic mechanisms related to climate events and their consequences.

EFFECTS OF SMALL PROBABILITIES

An important distinction has been made between uncertain or risky decisions made from personal experience and those made from statistical description, because the ostensibly same information about possible outcomes and their likelihood of choice alternatives can lead to different choices depending on how the information was acquired (Hertwig et al., 2004, 2006). Decisions from experience rely on (repeated) personal encounters with uncertain choice options, the way animals make risky foraging decisions (Weber, Shafir, and Blais, 2004). While the outcomes of choice options may initially be completely unknown, repeated choices provide the decision maker with feedback about possible outcomes and their likelihood, in the limit with great objective accuracy. Decisions from description, on the other hand, are made based on outcome and probability information provided in some statistical summary that is communicated in verbal, graphic, or numeric form. This way of information communication and acquisition is available only to humans, with their ability for abstract, symbolic representation, but is the method on which almost all laboratory studies of risky decision making have been based (Weber, Shafir, and Blais, 2004).

Knowing *how* people have come to know about the possible outcomes of real-world choice options and their probabilities matters, because choices differ quite dramatically under the two information conditions when choice options include small-probability events. Members of the general public and domain experts often learn about choice option outcomes and their probabilities in different ways. In the case of insurance decisions (e.g., federally subsidized flood insurance, Kunreuther, 1984), individuals considering the purchase of insurance appear to make their decision based on personal experience with flood events in previous years, whereas the industry experts have access to actuarial information and thus make decisions from description. In the case of childhood inoculation decisions, the situation is the opposite. The pediatrician who administers hundreds of inoculations per year knows about the outcomes and their probabilities of inoculating or failure to inoculate from personal experience, whereas parents make this decision based on a description of outcomes provided in medical-information pamphlets or on websites. Weber, Shafir, and Blais (2004) and Hertwig et al. (2006) described the association- and affect-based learning mechanisms by which personal experience with low-probability events leads to more apparent risk taking than that observed when the same options are presented by statistical summary descriptions. People's evaluations of risky options under repeated sampling follow classical reinforcement learning models where initial impressions are continuously updated in a way that gives recent events more weight than distant events.[1] Because rare events have a small(er) probability of having occurred recently, they (on average) tend to have a smaller impact on the decision than their objective likelihood of occurrence would warrant.[2] In those rare instances where they do occur, recency weighting gives them a much larger impact on the decision than warranted by their probability, making decisions from experience more volatile across respondents and past outcome histories than decisions from description. In contrast, the probability weighting function of prospect theory, which was developed to account for data sets that describe risky decisions from description, predicts that decision makers in decisions from description will overweight small probability events, that is, give them more weight in their decisions than they deserve based on their likelihood of occurrence.

Consistent with these predicted differences in the weight given to rare events under the two information conditions, people living in flood plains—who, as mentioned above, make decisions about flood insurance based on their personal experience with floods, a low base-rate event—have tended to turn down even federally subsidized insurance (Kunreuther, 1984), which is consistent with an *under* weighting of the actuarial frequency of such floods. Parents contemplating inoculations of their children against childhood diseases that have a low probability of life-threatening side effects, who make this decision based on statistical summary information about the benefits and side effects, have often turned down epidemiologically effective inoculations, which is consistent with an *over*-weighting of the low probability of severe side effects.

Slovic, Kunreuther, and White (1974) argued for the importance and utility of studying bounded rationality in field settings and already predicted that incremental personal experience of natural hazards and decisions based on such information may not be captured by expected utility models and their extensions. Recent work on important differences in decisions from experience and decisions from description confirm their predictions. The relative indifference with which most politicians and members of the general

public consider small-probability–high-consequence events like catastrophic rainfall and bursting levies, until and unless they have recently occurred, is much closer to the predictions made by the reinforcement learning models of decisions from experience than to that of prospect theory for decisions from description.

Useful Behavioral Insights

This section will review insights from behavioral decision research that may offer more hopeful predictions for the feasibility of environmentally responsible and sustainable decisions. A better appreciation of the three types of cognitive abundance that will be reviewed in this section will provide environmental policy makers and those charged with implementing such policies with tools to shape the decision environment in ways that may facilitate more environmentally sustainable behavior. In particular, I will show that there is utility in knowing that there are (1) multiple ways to represent (or frame) choice options that influence decisions, (2) multiple goals held by decision makers, typically in parallel, that are activated to different degrees by contextual features, and (3) multiple qualitatively different modes in which people can arrive at a decision, with the mode or process often influencing the outcome.

Multiple Representations: Framing and Mental Accounting

People have been found to represent choice options in different ways that, while normatively equivalent, nevertheless affect their decisions.

GAIN VERSUS LOSS FRAMING AND RISK AND LOSS AVERSION

Our neural system is set up in ways that makes the relative evaluation of outcomes much easier and more accurate than absolute evaluation (Weber, 2004). As a result, people can be expected to search for implicitly or explicitly provided reference points in the environment by which to judge the value of outcomes (Hsee, 1996). Shifting the perspectives of decision makers in ways that change their subjective evaluations of choice options is referred to as framing (Kahneman and Tversky, 1984). Often such changes in perspective are brought about by moving the decision maker's point of reference. Given what we know from prospect theory about differential risk attitudes for gains versus losses and about loss aversion, it is obvious that choice selection can be influenced dramatically if the up-front costs can be reframed not as losses but rather as forgone gains. In a simulation based on realistic farm cultivation decisions involving economic and physical conditions and crop models, Podesta et al. (2008) showed that changes in the reference point by which farmers encode farm profits as either gains or losses strongly affect what combination of crops turn out to be optimal, if farmers are assumed to attempt to optimize their returns as evaluated by a prospect-theory value function rather than by an expected-utility function. Another variable that differs quite significantly as a function of the reference point for returns (and thus the region of the return distribution that is encoded as a loss and subject to loss aversion) is the value of information (VOI) of available seasonal climate forecasts, which tell farmers probabilistically, but with some measure of skill, whether the coming growing season is of an El Niño, a La Niña, or a "normal" type. Whereas the VOI of such climate forecasts is on average positive, in the vicinity of 6%–7% (meaning that farmers' satisfaction with their returns can be expected to improve by this percentage if they use the climate forecast in an optimal fashion), for some combination of parameter values (high reference points or aspiration levels and large loss aversion), the VOI can actually be negative (Letson et al., 2009). These results suggest more generally that policy makers need to better understand decision makers' utility function and reference points in order to evaluate the impact of technological innovations and policy interventions.

SOCIAL COMPARISONS AND REGRET

The outcomes obtained by others provide a very salient reference point for relative comparisons. Regret theory, which was independently first proposed by Bell (1982) and Loomes and Sugden (1982), assumes not only that people make such comparisons after the fact (feeling somewhat good when they fared better than others, and very bad when they fared worse), but also that they anticipate these comparisons and incorporate them into their original decision of what to do. If regret about worse outcomes and rejoicing about better outcomes were of equal magnitude, anticipation of such emotions would cancel each other out. The assumption that regret is stronger than rejoicing puts regret theory into the class of models that assume that people often operate under asymmetric loss functions, where an error in one direction is seen as resulting in more severe consequences than an error in the opposite direction (Weber, 1994). The experience of strong regret following the mental comparison of a decision's unfavorable outcome with better outcomes that would have been obtained had a different decision been made has obvious teaching functions and

can improve the quality of decisions. The prominent use of available climate-change adaptation technologies by trusted opinion leaders (e.g., seasonal climate forecasts that help adaptation to the greater climate variability known to accompany climate change; or the use of more drought-resistant seed corn in agriculture) can be a way of putting experienced and subsequently anticipated regret about worse outcomes to work to help later adopters of such technologies modify their practices in a more timely fashion.

Advocacy of the precautionary principle to guide action in situations of highly uncertain but potentially very high stakes decisions can be traced back to a host of political and otherwise strategic motivations (Löfstedt, Fischhoff, and Fischhoff, 2002). Those motivations do not exclude, however, an intuitive psychological appeal of the principle, based on the anticipation of regret that could be extremely large (even if experienced only under a low-probability future state of the world) if human habitability of planet earth were to be compromised by the failure to take action because of the desire to not incur manageable economic costs.

DECISION-MAKING UNIT

Another way in which environmentally relevant decisions could be reframed in ways that might affect choices is by changing the focus of such decisions from individuals to groups. The decision makers' default foci of attention may be on themselves (i.e., on their needs, goals, and interests), since human processing limitations guide people into the direction of minimal effort, and personal needs, goals, and interests are most easily ascertained and most important. However, this typical attentional focus can be influenced by both the immediate decision environment and the more chronic surrounding cultural environment. Priming of broader social identities (e.g., national identity by a country's flag or other cultural icons) have long been used in times of war or other conflict to induce people to incur personal sacrifices for the sake of larger collectives and future times. Milch et al. (2009) showed that something as simple as the decision-making unit can focus attention on different goals and motivations. When groups of three people considered a delay-of-gratification decision (that affected them individually, as well as the group as a whole) for the first time collectively, they showed much greater patience and spent less time discounting than individuals either alone or in a group who had first considered the decision individually.

The "we" in a broader frame does not even need to be interpreted as "I and others." It can also refer to "my current self and my future self." Thus Bartels and Rips (2010) showed that individual differences in the perceived closeness to future selves had implications for how much people were willing to sacrifice current consumption for future consumption. When people make choices for their future selves, those choices tend to be affected more by rational, and less by affective, considerations and tend to resemble the choices they would make for other people (Pronin et al., 2008). In an extension, Wade-Benzoni (2009) found that people's perceived distance to future generations was inversely related to their generosity toward those generations.

Social psychological research has shown that group identity that turns the decision maker and actor from an "I" to a "we" can be induced by very minimal manipulations (Brewer, 1979). In any given decision, such changes in focus from individual identity and individual goals to group identity and group goals will be transient. However, as with any repeated execution of a process or behavior, initially transient and effortful processes become more chronic and automatic over time (Schneider and Chein, 2003). Cultures that emphasize the importance of affiliation and social goals over autonomy and individual goals have been shown to influence the way in which decisions under risk and uncertainty get made (Weber and Hsee, 1998), and different cultural emphases on individualism vs. collectivism are reflected in cultural products that shape chronic attention, from children's books, to proverbs and novels (Weber, Ames, and Blais, 2004; Weber and Hsee, 1998) and in cultural institutions and other affordances (Weber and Morris, 2010).

MENTAL ACCOUNTING

Mental accounting, or people's tendency to post financial and other income and expenses to separate accounts with different rules (Thaler, 1980), has often been depicted as a somewhat irrational adaptation to finite mental capacity and to self-control issues (Heath and Soll, 1996). However, the principle of psychological judo can be applied to this behavioral regularity as well, and somewhat dysfunctional behavior can be used at times as a tool that helps decision makers achieve their own best long-term interests. Just as new life events and attendant new goals result in the setting up of physical accounts (e.g., a savings account to pay for future college expenses of a new baby), goals can be made more salient to decision makers by helping them set up mental accounts for those goals. Concrete and vivid concepts like a person's carbon footprint, which can be measured using simple web tools (e.g., http://www.carbonfootprint.com/calculator.aspx), have played an important role in raising awareness among members of the general

public (in the Western world, at least) about their personal impact on local and global carbon dioxide levels. Whereas much remains to be done to make existing personal carbon-footprint calculators consistent and transparent (Padgett et al., 2007), these physical accounts facilitate the establishment of a mental account and, more importantly, provide a metric on which personal progress can be tracked. Setting up such accounts is especially effective if paired with actionable suggestions about how carbon dioxide emissions can be reduced with no financial penalties (Granade et al., 2009). Websites or personal consultants (e.g., http://www.carbon-partner.com/) who provide calculation aids to determine one's carbon footprint help individuals overcome attentional and information-processing limitations. Organizations that provide low-transaction-cost, web-based ways of offsetting carbon dioxide producing activities are an easy way for individuals to alleviate the guilt produced by an affective processing of the situation or to put their carbon dioxide account back into the black if the situation is processed analytically, though some have recently questioned whether these solutions are too "easy" (Wish, 2008) and may actually result in increased CO2 emissions, likening such offsets to modern indulgences.

Multiplicity and Flexibility of Goals

People's behavior is motivated by a broad range of goals, from individual goals of self-preservation and procreation; to more social goals, such as feeling connected; to meta-cognitive goals, such as feeling confident or in control. Various taxonomies of human needs—in sociology (Weber, 1921/1984), philosophy (Habermas, 1972), and psychology (Hilgard, 1987)—suggest that human needs are far broader than the maximization of personal material survival or genetic propagation. While material needs and instrumental goals (the human needs acknowledged for rational-economic man) are important, other classes of needs also play important roles. Social needs, for example, include both affiliation (wanting to belong) and individuation (asserting one's autonomy and uniqueness). Tyler (this volume) also emphasizes the fact that social motivation matters.

In any given situation, people have a multiplicity of goals. Choice options may be evaluated on their ability to satisfy the largest number of active goals, and new choice options may be generated if existing ones do not allow the decision maker to reach all of the important goals (Krantz and Kunreuther, 2007). To the extent that different goals in many situations are contradictory (e.g., wanting to consume *and* to conserve), decisions typically require a trade-off between the extent to which one or the other goal can be satisfied, even though people dislike this realization and have evolved ways of making decisions that minimize conscious tradeoffs (Payne, Bettman, and Johnson, 1993).

Specific human needs or goals can be temporarily activated by the nature of the choice set (Krantz and Kunreuther, 2007), primes in the external environment (North, Hargreaves, and McKendrick, 1997), or by a preceding task that implies the goal in question (e.g., a communal writing task, where a group of students compose a joint letter to the dean; Arora et al., 2009). North, Hargreaves, and McKendrick (1997) found that German wines were purchased 73% of the time when German (rather than French) background music was playing in the store, and French wines 77% of the time when French music was playing, even though customers were not aware of any effect on their purchase behavior. Arora et al. (2009) reported that cooperation in a social dilemma game went up from 43% to 75% when the preceding task required cooperation, relative to a control where individual students each previously had to act on their own.

Multiple Modes of Making Decisions

In the section "Behaviors of Concern," several pieces of evidence suggested that environmentally relevant decisions (e.g., choices in common-pool resource dilemmas) are seriously handicapped if people consider them in an analytic or calculation-based way, either with the unbounded rationality of selfish individual utility maximization or in a boundedly rational fashion (Marx and Weber, 2011). Fortunately, people are not restricted to making such decisions in an analytic way that compares costs and benefits and weighs outcomes and their probabilities. Calculation-based decision making may not even be the mode most prominently used by most people most of the time to make these and other decisions (Weber, Ames, and Blais, 2004). In this section, we further describe calculation-based, as well as other modes, of making decisions.

Weber and Lindemann (2007) distinguished between three classes of decision modes, namely calculation-based, affect-based, and recognition-based, which are referred to colloquially as decisions made by the head, by the heart, and by the book. These three classes of decision modes encode and utilize different situational inputs and apply different psychological processes. Calculation-based decisions involve analytical thought. Affect-based decisions are based on immediate, holistic, affective reactions (Damasio, 2000) and include impulse shopping (i.e., approach behavior that is driven by positive affect

toward the object of purchase) and decision avoidance (i.e., avoidance behavior that is driven by negative affect toward situations that offer no positive choice options or are too complex).

In recognition-based decision making, the decision maker recognizes a decision situation as a member of a class for which a satisfactory action is known (Simon, 1990). Recognition-based decisions come in different variants. In case-based decisions, the decision maker is typically an expert with a memory store full of specific situations in her domain of expertise, with the most appropriate action stored for each one. These mental representations can be thought of as "if-then" productions, where the *if* element is a set of conditions that must be met in order to trigger the resultant action represented by the *then* part of the production. The expert decision maker is able to unconsciously apply these production rules, which have been developed through repeated experience, as has been suggested by research on experts such as firefighters and jet pilots (Klein, 1999).

Rule-based decisions are another type of recognition-based decisions. These rules may be laws ("if you are driving and come to a red light, then you must stop") or other types of regulations (parental rules, self-imposed admonishments, societal norms, or company rules) (Prelec and Herrnstein, 1991). In role-based decisions, the decision context elicits a rule of conduct that derives from one of the social roles of the decision maker (March and Heath, 1994). Roles can include positions of responsibility within society (the role of parent), group memberships (the role of being a Christian), and self-defining characteristics (the role of being honest). Each of these roles has associated obligations that are recalled and executed when a triggering situation is encountered.

Implicit rules and role-related obligations are often acquired through observational learning and imitation. Sociologists and psychologists (from Ellwood [1901] to Sloate and Voyat [2005]) have long argued that modern notions of the autonomous self have falsely emphasized the role of individual decisions on human behavior over that of social influences. Copying the observed behavior of others is a widespread phenomenon of which the imitator is typically unaware and plays a large role in human development (Meltzoff and Moore, 1999).

Conditioned responses and habits acquired without conscious awareness probably determine a large amount of behavior. Unconscious processing occurs at the encoding stage of learning, where much information is stored for future use without our explicit intention (Reber, 1996), and at the retrieval stage, where primes in the external or internal environment can increase the accessibility of a subset of information, goals, or intentions, thus influencing observed behavior (Weber and Johnson, 2006). For example, the dimension of comfort versus price could be primed by exposing internet shoppers who were looking for sofas to either feathery clouds or $ symbols, respectively, in the background wallpaper of the initial web store page. Shoppers bought sofas that scored higher on the primed dimension (Mandel and Johnson, 2002). Emotional reactions above or below our level of awareness also often mediate learning by leading to approach and avoidance responses (Damasio, 1994; Loewenstein et al., 2001). While conscious learning or problem solving typically requires that the individual perceives there to be a problem, learning or adaptation without awareness has the advantage that it can happen without a conscious diagnosis that something is wrong and requires action.

Different decision modes can be executed in parallel and differ in their time course, with the more automatic ones turning in their verdict earlier, while the more conscious and effortful ones require more time to completion. People report using between two or three modes for any given decision (Krosch, Figner, and Weber, 2009; Weber and Lindemann, 2007). When the choice option selected by different decision modes is the same, cross-modal consensus on the best action contributes to decision confidence. When the indicated best choice option differs between decision modes, the relative weight given to the output of the different modes will determine which one gets selected, and decision confidence will be low(er) (Weber et al., 2000).

Engel and Weber (2007) discuss how the human information-processing system might implicitly decide which mode of decision making to apply in a given situation, or to which decision mode to give the deciding weight in situations of choice conflict. High-level goals or motives with high activation levels in a particular cultural context have been shown to influence the choice of decision modes. When cultures differentially emphasize the importance and desirability of such goals as autonomy versus social connectedness, for example, different decision modes become more prevalent, because different modes are differentially suited to satisfy those goals. Thus Weber, Ames, and Blais (2004) found that characters in Chinese twentieth-century novels, who operated in a collectivist culture emphasizing affiliation, were more likely to make role-based decisions and less likely to make affect-based decisions than characters in American twentieth-century novels, who operated in an individualist culture with its emphasis on autonomy. Western, consumption- and economic-growth oriented societies and their formal and informal institutions (including education, advertising,

entertainment, and the media) may be priming values and goals that are incompatible with more environmentally sustainable behavior. Their conceptualization of progress through competition, both against other economic or political players and/or against oneself over time, may stack the cards against the resource conservation and cooperation needed to overcome common-pool resource dilemmas, unless such competition can be redirected toward (friendly) competition to achieve carbon dioxide mitigation and other sustainability goals.

This previous discussion suggests that policy interventions should be designed to prime social roles that will induce people to use rule-based processes to determine their environmentally relevant behavior, which may necessitate changes in the dominant culture and its primes in Western countries.

Discussion

The goal of environmental policy is to change behavior of companies, governing boards and committees, and members of the general public in the direction of more sustainable, long-term, and socially and environmentally responsible actions. Conventional policy interventions do so either by command and control or by changing incentives, applying both carrots to encourage desirable behavior and sticks to discourage undesirable actions. This chapter has argued that this understanding of policy intervention options is too narrow in at least two ways. First, conventional policy interventions are not using the full range of goals that motivate behavior and changes in behavior. Second, conventional policy interventions do not utilize the full range of processes that people use to decide on a course of action. The tools described in the previous section (multiple and flexible goals, representations, and decision modes) suggest that there might be cheaper and more effective ways of achieving environmental goals than taxes and regulations. Evidence that neither the public nor markets are fully responsive to material incentives comes from the fact, already mentioned earlier, that many existing energy-efficiency increasing technological innovations (e.g., CFL or LED lighting, space heating and cooling technology) are nowhere near fully utilized. This is despite the fact that changing to such technologies constitutes net present-value savings, that is, the initial expenditures to switch are more than fully compensated by subsequent energy-cost savings over the lifetime of the devices. Why do consumers resist change even for such "low-hanging fruit" that provide net energy-cost savings, without any compromises in lifestyle and with positive social and environmental effects, let alone changes in environmentally relevant behavior that are perceived as requiring upfront sacrifices in quality of life? Are there ways in which policy makers can reframe such choices in ways that decision makers will act in individually, socially, and environmentally more beneficial ways willingly?

If people were rational-economic decision makers, the answer to this question would be the provision of better information about the possible risks of status-quo behavior to themselves or their children or grandchildren and about the benefits of changes in behavior. One obstacle to the success of rational calculation-based approaches in bringing about environmentally responsible, that is, sustainable-growth-promoting, behavior, even if people were Bayesian updaters and utility maximizers, is the fact that most of the costs and benefits of different behavioral options lie well into the future, with the result that the relative expected utility of different options depends critically and almost exclusively on one factor, namely the rate at which people discount future outcomes (Weitzman, 2007). This effectively turns calculation-based decision making in this domain into a philosophical or ethical question about the "correct" discount rate to use.

A large and growing literature on behavioral, and in particular psychological, issues in discounting (Loewenstein and Elster, 1992; Weber et al., 2007) has equal, and perhaps more important, implications for policy design. While there is some evidence to suggest that people discount outcomes in different domains differently (e.g., are even more impatient for immediate positive health outcomes than they are for immediate financial outcomes [Chapman, 1996]), environmental and financial outcomes seem to be discounted at very similar rates (Hardisty and Weber, 2009). In addition, domain differences in discounting are much smaller than differences in the discounting of future outcomes when they are being framed as either gains or losses, with much less discounting of future losses (Hardisty and Weber, 2009). These and other behavioral results suggest that there are different psychological, economic, and ethical reasons for discounting, which need to be better understood and disentangled (Hardisty et al., 2010), because they have different implications for policy design.

If not absence of information in the face of optimal or even biased cost-benefit calculations, what other reasons contribute to people's reluctance to change their behavior in environmental, as well as in other, contexts? This chapter points to automatic processes and behavior as contributing causes of people's apparent status quo bias. This suggests that people need to be jolted into changing any thoughtless resource-consuming habits, perhaps by scaring them, in the way movies like *The Day after Tomorrow* or

An Inconvenient Truth have been trying to do for climate change mitigation. Unfortunately, as described in the first part, people do not appear to be easily or lastingly scared by chronic environmental risks, and appeals to fear have important drawbacks in general.

The second part of this chapter provided some more positive suggestions about how to attract people's attention to environmental risks, that is, how to generate the *ex ante* concern that these risks seem to warrant *ex post*. The concretization of future events and movement of them closer in time and space seem to hold promise as interventions that will raise visceral concern (Marx et al., 2007). For example, simulations that provide visual evidence of the projected 10–30 year effects of plausible sea-level rises in people's favorite summer vacation location or of the disappearance of snow covers in their favorite ski resort may well raise visceral concern. Such interventions would need to be conducted with full awareness about unintended side effects (like reductions in concern about other important risks) and in ways designed to help people overcome cognitive and affective capacity limitations (e.g., the single-action bias).

Query theory (Johnson, Häubl, and Keinan, 2007; Weber et al., 2007) suggests that guided protocols by which decision makers consider arguments for conservation and climate change mitigation before they are allowed to consider arguments for staying with the status quo will help to improve the balance of support between the desire for immediate gratification and the goal of longer-term environmental preservation or sustainability toward the latter. Finally, better education in (environmental) science and statistics can create the familiarity with the scientific presentation of information that will reduce people's aversion to behavior options with uncertain consequences and may create citizens who give greater weight to the less-distorted output of their analytic processing system, moving the risk perception of the general public and its officials closer to that of environmental scientists.

Rule-based decision making, where choice follows from consciously or more automatically triggered norms of conduct, also seems to offer considerable advantages. Evangelical churches and other Christian denominations in the United States have recently started to reframe the debate about economic development versus environmental conservation from one about material costs and benefits to one of moral responsibilities and obligations (see Robinson and Chatraw, 2006). When a message about the moral imperative to preserve our planet with its natural resources for future generations ("stewardship of the earth") comes from credible sources (e.g., the National Council of Churches in the United States), decisions about consumption or conservation of resources will less likely be made by personal and myopic cost-benefit calculations and more likely by role- and rule-based decision processes that are less susceptible to impatience and excessive discounting of future consequences. There probably is considerable cultural universality in the effectiveness of such reframing of consumption decisions from a calculation of costs and benefits to one of responsibilities and obligations. What can be expected to differ across countries or cultures is the most effective organization to issue or endorse the norms of responsible consumption and stewardship. While evangelical or Christian churches may be a natural source of such rules of behavior in the United States, in more secular countries this role could fall to political organizations (e.g., green parties in Europe). In general, institutions with a long time horizon and a nongeographic and nonnationalistic focus that possess the trust of the general population would seem to be ideal issuers and disseminators of a philosophy and ethic of resource stewardship, preservation, and responsible and mindful consumption.

Even with the best intentions (e.g., about responsible stewardship of the earth), the devil is in the details. Goals and attitudes do not always translate into intended action (Gollwitzer, 1999). Attention to one's goals waxes and wanes with the activation levels of these goals, and many consumption behaviors have become partially or fully automatic, that is, they happen without much conscious thought. Overcoming these habitual behaviors may require explicit coaching followed by constant reminders and frequent feedback. Such active interventions are effortful, both for the mentor and the mentee. With humans' finite attention and processing capacity, more passive approaches toward keeping attention on the relevant goals and on instantiating the desired behaviors have much to commend themselves.

Measuring the costs of thoughtless consumption behavior by prominently displayed meters that provide constant feedback could be one way to draw and keep people's attention on the goals of saving and conservation. The arrival of smart-grid technology will enable careful experimentation with the best way of putting metering and feedback devices into action without overstretching people's processing capacity or losing their attention over extended periods of time. Immediate and prominent feedback (e.g., online fuel-efficiency calculation for a car, in a prominent dashboard display) can turn conservation into a video game where players can improve on their own previous record and can also engage in friendly competitions against other players on websites where their accomplishments can be listed.

Socially desirable goals can be kept active, either chronically or at strategic moments of important decisions, by designing decision environments that expose people to reminders of these goals or by designing

social environments that prime these goals and thus keep them active (Weber and Morris, 2010). The fact that the economic development of countries or regions is related to the degree of civic engagement of its population (Putnam, 1995), for example, can be explained by the greater chronic accessibility of economic-development-enhancing goals (which are related to social capital) as the result of recreational activities that require and foster these goals.

Another promising avenue toward encouraging environmentally responsible behavior in a low-effort and low-awareness way is the use of social influence, observational learning, and imitation. People are influenced by the behavior of others even in such seemingly rational settings as financial markets, and social influence seems to be particularly prevalent in situations with ambiguity about the best way to proceed (Schoenberg, 2007). Imitation can lead to behavioral change without any need for the realization that change is in order. At levels beyond the individual, demonstration projects conducted by visible groups or companies can serve a similar function, not only by showing the feasibility of a particular new technology or institution but also by triggering imitation on the part of other players.

A final promising tool is the judicious use of decision defaults (see Johnson and Goldstein, this volume). Most decisions have explicit defaults (e.g., nonresponse to a letter from a utility company will result in continuation of the electricity being provided from nonrenewable sources) even when these are not clearly spelled out. Only very rarely do we encounter situations where an active decision must be made. Given that defaults are unavoidable and do not diminish people's ability to choose the option they truly prefer if they are willing and able to process all available information to make an informed calculation-based decision, behavioral economists like Benartzi and Thaler (2004, this volume) argued, in the context of supplementary pension-savings decisions, that decision defaults should be set to the choice option that maximizes people's own long-term utility rather than to an option that the decision maker will later regret taking. Setting judicious decision defaults will ensure that people are not hurt by decision avoidance that might be triggered by informational overload and lack of interest (Goldstein et al., 2008). A similar argument has been made for decisions that impact other individuals and society at large, for example, organ donation. Johnson and Goldstein (2003) showed that differences in the decision default (from "opting in," that is, not being a donor unless one decides to be one to "opting out," that is, being a donor unless one decides not to) in European countries led to stunning differences in stated willingness to donate as well as in actual organ transplantations, with effects that far exceeded any other interventions, including very expensive information campaigns. Such observations have implications for environmentally relevant domains such as building codes, where energy-efficient and environment-friendly choice options should and could appear as decision defaults, thus greatly increasing their uptake (Dinner et al., 2009).

Query theory (Johnson et al., 2007; Weber et al., 2007) correctly predicts people's failures of hedonic forecasting in the case of changes in status quo. Changes to options other than the status quo are often resisted because people tend first to generate arguments for the status quo options, and subsequent queries that explore arguments for other choices are disadvantaged because of output interference. What people fail to realize is the fact that this process also kicks in after a default has been changed, often against their better judgment, so that their future evaluations of the new default tend to be far more positive *ex post* than they were *ex ante*. One such example is the smoking ban in bars imposed in New York City by Mayor Bloomberg in 2006 against much *ex ante* industry and public opposition, which was evaluated quite positively *ex post* by a large majority of New Yorkers a year or two later.[3] Query theory and such examples suggest that policy makers may sometimes be well advised to shape and lead public opinion rather than follow it.

Failing these various efforts to help individuals overcome their egocentric and present-focused myopia and lack of hardwired affective early-warning responses to environmental concerns that typically require present actions to prevent future problems, many environmental problems can be expected to increase in both the severity and detectability of negative consequences. While mounting personal and local evidence of such phenomena as climate change and its potentially devastating consequences can be counted on to be an extremely effective teacher and motivator in future years, these lessons may unfortunately arrive too late for corrective action.

Notes

Preparation of this chapter was facilitated by a residential fellowship at the Russell Sage Foundation in 2007/08 and by two grants from the National Science Foundation (SES-0345840 and SES-0720452) and a grant by the National Oceanographic and Atmospheric Administration.

1. This sort of updating and learning is adaptive in dynamic environments, where circumstances might change with the seasons or according to some other cycles or trends.

2. An additional reason that rare events get underweighted is that with small samples, they often are not experienced at all and hence do not enter into the decision at all. The underweighting of small-probability events does not

depend on just these cases, however, but follows from the iterative updating rule, where the most recent event gets a high weight and the weight of previous events decays fairly rapidly. Rare events get underweighted on average, because they have a small(er) probability of occurring as the most recent event than more likely events.

3. I thank Eric Johnson for this example.

References

Ainslie, G. (1975). Specious reward: A behavioral theory of impulsiveness and impulse control. *Psychological Bulletin*, *82*, 463–496.

Allais, P. M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école americaine. *Econometrica*, *21*, 503–546.

Ames, D. R., Flynn, F. J., and Weber, E. U. (2004). It's the thought that counts: On perceiving how helpers decide to lend a hand. *Personality and Social Psychology Bulletin*, *30*, 461–474.

Arora, P., Peterson, N. D., Krantz, D. H., Hardisty, D. J., and Reddy, K. R. (2009). *Testing the limits of group affiliation on cooperation in social dilemmas*. Working paper. Center for Research on Environmental Decisions (CRED).

Bartels, D. M., and Rips, L. J. (2010). Psychological connectedness and intertemporal choice. *Journal of Experimental Psychology: General*, *139*, 49–69.

Bell, D. E. (1982). Regret in decision making under uncertainty. *Operations Research*, *30*, 961–981.

Benartzi, S., and Thaler, R. H. (1995). Myopic loss aversion and the equity premium puzzle. *Quarterly Journal of Economics*, *110*, 73–92.

———. (2004). Save more tomorrow: Using behavioral economics to increase employee saving. *Journal of Political Economy*, *112*, 164–187.

Bernoulli, D. (1738/1954). Exposition of a new theory on the measurement of risk (L. Sommer, Trans.). *Econometrica*, *22*, 23–36.

Bowles, S. (2004). *Microeconomics: Behavior, institutions, and evolution*. Princeton, NJ: Princeton University Press.

Brewer, M. B. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin*, *86*, 307–324.

Camerer, C. (2000). Prospect theory in the wild. In D. Kahneman, and A. Tversky (Eds.), *Choice, values, and frames* (pp. 288–300). New York: Cambridge University Press.

Chaiken, S., and Trope, Y. (1999). *Dual-process theories in social psychology*. New York: Guilford.

Chapman, G. B. (1996). Temporal discounting and utility for health and money. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 771–791.

Damasio, A. (1994). *Descartes' error: Emotion, reason and the human brain*. New York: Penguin Books.

———. (2000). *The feeling of what happens: Body and emotion in the making of consciousness*. San Diego: Harvest.

Dinner, I. M., Johnson, E. J., Goldstein, D. G., and Liu, K. (2009). *Partitioning default effects: Why people choose not to choose*. Social Science Research Network Working Paper. Retrieved from http://ssrn.com/abstract=1352488

Dunlap, R. E., and Saad, L. (2001, April 16). *Only one in four Americans are anxious about the environment*. Princeton, NJ: Gallup Organization. Retrieved from http://www.gallup.com/poll/1801/only-one-four-americans-anxious-about-environment.aspx

Ellsberg, D. (1961). Risk, ambiguity and the Savage axioms. *Quarterly Journal of Economics*, *75*, 643–669.

Ellwood, C.A. (1901). The theory of imitation in social psychology. *American Journal of Sociology*, 6, 721–741.

Engel, C., and Weber, E. U. (2007). The impact of institutions on the decision of how to decide. *Journal of Institutional Economics*, *3*, 323–349.

Evans, J.S.B.T. (2007). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255–278.

Fischhoff, B., Slovic, P., Lichtenstein, S., Read, S., and Combs, B. (1978). How safe is safe enough? A psychometric study of attitudes towards technological risks and benefits. *Policy Sciences*, *9*, 127–152.

Gillingham, K., Newell, R. G., and Palmer, K. (2009). Energy efficiency economics and policies. Discussion Paper dp-09-13. Washington, DC: Resources For the Future. Retrieved from http://www.rff.org/documents/RFF-DP-09-13.pdf

Goldstein, D. G., Johnson, E. J., Herrmann, A., and Heitmann, M. (2008, December). Nudge your customers toward better choices. *Harvard Business Review*, pp. 99–105.

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist*, *54*, 493–503.

Granade. H. C., Creyts, J., Derkach, A., Farese, P., Nyquist, S., and Ostrowski, K. (2009). *Unlocking energy efficiency in the U.S. economy*. Technical report. McKinsey Global Energy and Materials, McKinsey and Company.

Habermas, J. (1972). *Knowledge and human interests* (Trans. J. Shapiro). London: Heinemann.

Hansen, J., Marx, S., and Weber, E. U. (2004). *The Role of climate perceptions, expectations, and forecasts in farmer decision making: The Argentine Pampas and South Florida*. Technical Report 04-01. Palisades, NY: International Research Institute for Climate Prediction (IRI).

Hardin, G. (1968). The tragedy of the commons. *Science*, *162*, 1243–1248.

Hardisty, D. J., Johnson, E. J., and Weber, E. U. (2010). A dirty word or a dirty world? Attribute framing, political affiliation, and query theory. *Psychological Science*, *21*, 86–92.

Hardisty, D. J., Orlove, B., Krantz, D. H., Small, A., and Milch, K. (2009). *About time: An integrative approach to effective policy*. Working Paper. Columbia Center for Research on Environmental Decisions.

Hardisty, D. J., and Weber, E. U. (2009). Temporal discounting of environmental outcomes: Effects of valence outweigh domain differences. *Journal of Experimental Psychology: General*, *3*, 329–340.

Heath, C., and Soll, J. B. (1996). Mental accounting and consumer decisions. *Journal of Consumer Research*, *23*, 40–52.

Heath, C., and Tversky, A. (1991). Preference and belief: Ambiguity and competence. *Journal of Risk and Uncertainty*, *4*, 5–28.

Hersch, J., and Viscusi, W. K. (2006). The generational divide in support for environmental policies: European evidence. *Climatic Change*, *77*, 121–136.

Hertwig, R., Barron, G., Weber, E. U., and Erev, I. (2004). Decisions from experience and the effect of rare events. *Psychological Science*, *15*, 534–539.

———. (2006). Rare risky prospects: Different when valued through a window of sampled experiences. In K. Fiedler and P. Juslin (Eds.), *Information sampling as a key to understanding adaptive cognition in an uncertain environment* (pp. 72–91). New York: Cambridge University Press.

Hilgard, E. R. (1987). *Psychology in America: A historical survey*. San Diego: Harcourt.

Holtgrave, D., and Weber, E. U. (1993). Dimensions of risk perception for financial and health risks. *Risk Analysis*, *13*, 553–558.

Hsee, C. K. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes*, *67*, 247–257.

Hsu, M., Bhatt, M., Adolphs, R., Tranel, D., and Camerer, C. F. (2005). Neural systems responding to degrees of uncertainty in human decision-making. *Science*, *9*, 1680–1683.

Johnson, E. J., and Goldstein, D. (2003). Do defaults save lives? *Science*, *302*, 1338–1339.

Johnson, E. J., Häubl, G., and Keinan, A. (2007). Aspects of endowment: A query theory of loss aversion. *Journal of Experimental Psychology: Learning Memory and Cognition*, *33*(3), 461–474.

Kahneman, D. (2003). A psychological perspective on economics. *American Economic Review*, *93*(2), 162–168.

Kahneman, D., and A. Tversky (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–292.

———. (1984). Choices, values, and frames. *American Psychologist*, *39*(4), 341–350.

Klein, G. (1999). *Sources of power: How people make decisions*. Boston: MIT Press.

Krantz, D. H., and Kunreuther, H. C. (2007). Goals and plans in decision making. *Judgment and Decision Making*, *2*, 137–168.

Krosch, A., Figner, B., and Weber, E. U. (2009). *Choice processes and their consequences in morally conflicting military decisions*. Working Paper. Columbia Center for Decision Sciences.

Kunreuther, H. (1984). Causes of underinsurance against natural disasters. *Geneva Papers on Risk and Insurance*, *31*, 206–20.

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics*, *112*, 443–477.

Letson, D., Laciana, C. E., Bert, F., Weber, E. U., Katz, R. W., Gonzalez, X. I., and Podesta, G. (2009). Value of perfect ENSO phase predictions for agriculture: Evaluating the impact of land tenure and decision objectives. *Climatic Change*, *97*, 145–170.

Levin, I. P., and Gaeth, G. J. (1988). Framing of attribute information before and after consuming the product. *Journal of Consumer Research*, *15*, 374–378.

Lichtenstein, S., and Slovic, P. (Eds.), (2006). *The construction of preference*. New York: Cambridge University Press.

Loewenstein, G., and Elster, J. (Eds.). (1992). *Choice over time*. New York: Russell Sage.

Loewenstein, G. F., Weber, E. U., Hsee, C. K., and Welch, E. (2001). Risk as feelings. *Psychological Bulletin*, *127*, 267–286.

Löfstedt, R., Fischhoff, B., and Fischhoff, I. (2002). Precautionary principles: General definitions and specific applications to genetically modified organisms (GMOs). *Journal of Policy Analysis and Management*, *21*, 381–407.

Loomes, G., and Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *Economic Journal*, *92*, 805–824.

Mandel, N., and Johnson, E. J. (2002). When web pages influence choice: Effects of visual primes on experts and novices. *Journal of Consumer Research*, *29*, 235–245.

March, J. C., and Heath, C. (1994). *Primer on decision making: How decisions happen*. New York: Free Press.

Marx, S., and Weber, E. U. (2011). Decision making under climate uncertainty: The power of understanding judgment and decision processes. In T. Dietz and D. C. Bidwell (Eds.), *Climate change in the Great Lakes region: Navigating an uncertain future*. East Lansing, MI: Michigan State Univ. Press.

Marx, S. M., Weber, E. U., Orlove, B. S., Leiserowitz, A., Krantz, D. H., Roncoli, C., and Phillips, J. (2007). Communication and mental processes: Experiential

and analytic processing of uncertain climate information. *Global Environmental Change*, *17*, 47–58.

McFadden, D. (1999). Rationality for economists? *Journal of Risk and Uncertainty*, *19*, 73–105.

Meltzoff, M. A., and Moore, M. K. (1999). Persons and representation: Why infant imitation is important for theories of human development. In J. Nadel and G. Butterworth (Eds.), *Imitation in infancy*. Cambridge: Cambridge University Press.

Milch, K. F., Appelt, K. C., Weber, E. U., Handgraaf, M.J.J., and Krantz, D. (2009). From individual preference construction to group decisions: Framing effects and group processes. *Organizational Behavior and Human Decision Processes*, *108*, 242–255.

Mischel, W., Grusec, J., and Masters, J. C. (1969). Effects of expected delay time on the subjective value of rewards and punishments. *Journal of Personality and Social Psychology*, *11*, 363–373.

North, A. C., Hargreaves, D. J., and McKendrick, J. (1997). In-store music affects product choice. *Nature*, *390*, 132.

O'Donoghue, T., and Rabin, M. (1999). Doing it now or later. *American Economic Review*, *89*, 103–124.

Ostrom, E., Dietz, T., Dolsak, N., Stern P. C., Stonich, S., and Weber, E. U. (Eds.) (2002). *The drama of the commons*. Washington, DC: National Academy Press.

Padgett, J. P., Steinemann, A. C., Clarke, J. H., and Vandenbergh, M. P. (2007). A comparison of carbon calculators. *Environmental Impact Assessment Review*, *28*, 106–115.

Payne, J. W., Bettman, J. R. and Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge: Cambridge University Press.

Peters, E., and Slovic, P. (2000). The springs of action: Affective and analytical information processing in choice. *Personality and Social Psychology Bulletin*, *26*, 1465–1475.

Pew Research Center. (2009). *Fewer Americans see solid evidence of global warming. Modest support for "cap and trade" policy*. Pew Research Center. Retrieved from http://www.people-press.org/2009/10/22/fewer-americans-see-solid-evidence-of-global-warming/

Podestá, G., Weber , E. U., Laciana, C., Bert, F., and Letson, D. (2008). Agricultural decision-making in the Argentine Pampas: Modeling the interaction between uncertain and complex environments and heterogeneous and complex decision makers. In T. Kugler, J. C. Smith, T. Connolly, and Y.-J. Son (Eds.), *Decision modeling and behavior in uncertain and complex environments* (pp. 57–76). Berlin: Springer.

Prelec, D., and R. J. Herrnstein. Preferences and principles: Alternative guidelines for choice. In R. Zeckhauser (Ed.), *Strategic reflections on human behavior* Cambridge, MA: MIT Press.

Pronin, E., Olivola, C.Y., and Kennedy, K. A. (2008). Doing unto future selves as you would do unto others: Psychological distance and decision making. *Personality and Social Psychology Bulletin*, *34*, 224–236.

Putnam, R. D. (1995). Bowling alone: America's declining social capital. *Journal of Democracy*, *6*, 65–78.

Read, D., Loewenstein, G., and Rabin, M. (1999). Choice bracketing. *Journal of Risk and Uncertainty*, *19*, 171–197.

Reber, A. S. (1996). *Implicit learning and tacit knowledge: An essay on the cognitive unconscious*. Oxford: Oxford University Press.

Robinson, T., and Chatraw, J. (2006). *Saving God's green earth: Rediscovering the church's responsibility to environmental stewardship*. Norcross, GA: Ampelon.

Schneider, W., and Chein, J. M. (2003). Controlled and automatic processing: Behavior, theory, and biological mechanisms, *Cognitive Science*, *27*, 525–559.

Schoenberg, E. (2007). *Beauty is in the eye of the other beholders: Strategy matching in financial markets* (Doctoral dissertation). Columbia University, New York.

Simon, H. A. (1982). *Models of bounded rationality* (Vols. 1 and 2). Cambridge, MA: MIT Press.

———. (1990). Invariants of human behavior. *Annual Review of Psychology*, *41*, 1–19.

Sloate, P. L., and Voyat, G. (2005). Language and imitation in development. *Journal of Psycholinguistic Research*, *12*, 199–222.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22.

Slovic, P. (1997). Trust, emotion, sex, politics, and science: Surveying the risk-assessment battlefield. In M. Bazerman, D. Messick, A. Tenbrunsel, and K. Wade-Benzoni (Eds.), *Psychological perspectives to environmental and ethical issues in management* (pp. 277–313). San Francisco: Jossey-Bass.

Slovic, P., Kunreuther, H., and White, G. F. (1974). *Decision processes, rationality and adjustment to natural hazards*. In G. F. White (Ed.), Natural hazards: local, national and global (pp. 187–205). New York: Oxford University.

Slovic, P., Lichtenstein, S., and Fischhoff, B. (1984). Modeling the societal impact of fatal accidents. *Management Science*, *30*, 464–474.

Sunstein, C. R. (2006). The availability heuristic, intuitive cost-benefit analysis, and climate change. *Climatic Change*, *77*, 195–210.

Thaler, R. H. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior and Organization*, *1*, 39–60.

Thaler, R., and Johnson, E. J. (1990). Gambling with the house money and trying to break even: The effects of prior outcomes on risky choice. *Management Science*, *36*, 643–660.

Trope, Y., and Liberman, N. (2003). Temporal construal. *Psychological Review*, *110*, 403–421.

Tversky, A., and Kahneman, D. (1992). Advances in

prospect theory, cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5, 297–323.

von Neumann, J., and Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

Wade-Benzoni, K. A. (2009). The egoism and altruism of intergenerational behavior. *Personality and Social Psychology Review*, *13*, 165–193.

Weber, E. U. (1994). From subjective probabilities to decision weights: The effect of asymmetric loss functions on the evaluation of uncertain outcomes and events. *Psychological Bulletin*, *115*(2), 228–242.

———. (1997). Perception and expectation of climate change: Precondition for economic and technological adaptation. In M. Bazerman, D. Messick, A. Tenbrunsel, and K. Wade-Benzoni. (Eds.), *Psychological perspectives to environmental and ethical issues in management* (pp. 314–341). San Francisco, CA: Jossey-Bass.

———. (2004). Perception matters: Psychophysics for economists. In J. Carrillo and I. Brocas (Eds.), *Psychology and economics* (pp. 165–176). Oxford: Oxford University Press.

———. (2006). Experience-based and description-based perceptions of long-term risk: Why global warming does not scare us (yet). *Climatic Change*, *70*, 103–120.

Weber, E. U., Ames, D. R., and Blais, A.-R. (2004). "How do I choose thee? Let me count the ways": A textual analysis of similarities and differences in modes of decision-making in China and the United States. *Management and Organization Review*, 1, 87–118.

Weber, E. U., Böckenholt, U., Hilton, D. J., and Wallace, B. (2000). Confidence judgments as expressions of experienced decision conflict. *Risk Decision and Policy*, *5*, 1–32.

Weber, E. U., and Hsee, C. (1998). Cross-cultural differences in risk perception, but cross-cultural similarities in attitudes towards perceived risk. *Management Science*, *44*, 1205–1217.

Weber, E. U., and Johnson, E. J. (2006). Constructing preferences from memory. In S. Lichtenstein and P. Slovic (Eds.), *The construction of preference* (pp. 397–410). New York: Cambridge University Press.

———. (2009). Mindful judgment and decision making. *Annual Review of Psychology*, *60*, 53–86.

Weber, E. U., Johnson, E. J., Milch, K., Chang, H., Brodscholl, J., and Goldstein, D. (2007). Asymmetric discounting in intertemporal choice: A query theory account. *Psychological Science*, *18*, 516–523.

Weber, E. U., and Lindemann, P. G. (2007). From intuition to analysis: Making decisions with our head, our heart, or by the book. In H. Plessner, C. Betsch, and T. Betsch (Eds.), *Intuition in judgment and decision making* (pp. 191–208). Mahwah, NJ: Lawrence Erlbaum.

Weber, E. U., and Morris, M. W. (2010). Cultural differences in judgment and decision making: Insights from constructivist, structuralist approaches. *Perspectives on Psychological Science*, *5*, 410–419.

Weber, E. U., Shafir, S., and Blais, A.-R. (2004). Predicting risk-sensitivity in humans and lower animals: Risk as variance or coefficient of variation. *Psychological Review*, *111*, 430–445.

Weber, E. U., Siebenmorgen, N., and Weber, M. (2005). Communicating asset risk: How name recognition and the format of historic volatility information affect risk perception and investment decisions. *Risk Analysis*, *25*, 597–609.

Weber, M. (1921/1984). *Soziologische Grundbegriffe*. Tuebingen, Germany: J.C.B. Mohr.

Weitzman, M. L. (2007). A review of the Stern review on the economics of climate change. *Journal of Economic Literature*, *45*, 686–702.

Wilde, G.J.S. (1998). The theory of risk homeostasis: Implications for safety and health. *Injury Prevention*, *4*, 89–91.

Wish, V. (2008). Pay as you go: The pros and cons of carbon offsets. Retrieved from http://knowledge.allianz.com/climate/?280/carbon-offsets-pros-cons

Woodward, R. T. (1998). *Should agricultural and resource economists care that the subjective expected utility hypothesis is false?* Selected paper presented at the meeting of the American Agricultural Economics Association. Salt Lake City, Utah. Retrieved from http://ageconsearch.umn.edu/bitstream/20941/1/spwood01.pdf

CHAPTER 23

Overcoming Decision Biases to Reduce Losses from Natural Catastrophes

HOWARD KUNREUTHER

ROBERT MEYER

ERWANN MICHEL-KERJAN

In May 2008 a storm surge triggered by Cyclone Nargis swept through low-lying coastal areas of Myanmar without warning, killing an estimated 138,000 residents (Fritz et al., 2009). As staggering as this loss of life was, it was dwarfed by the estimated 230,000 who had died four years earlier in eleven countries in Southeast Asia from a major earthquake and accompanying tsunami (http://www.tsunami2004.net). Even wealthy countries that have the resources to invest in risk-reducing (i.e., mitigation) measures and warning systems have recently witnessed significant losses from natural hazards, such as the $150 billion in total economic damages and 1,300 deaths caused by Hurricane Katrina in Louisiana and Mississippi in 2005 (Knabb et al., 2006). Perhaps most disturbing, if these events taught lessons about the potency of natural hazards, they were not widely absorbed by those at risk. Just three years after Katrina, many residents of the Bolivar Peninsula in Texas refused to heed urgent evacuation warnings as Hurricane Ike approached, a reluctance that led to the deaths of over 100 and forced a major airlift of survivors after the storm (Berg, 2009).

While these events provide perhaps the most vivid examples of recent disasters, they are but part of a trend of escalating losses from natural hazards that has been observed over the past twenty years. If one considers the twenty-five most costly insured catastrophes anywhere in the world between 1970 and 2008, all of them occurred after 1987. (Kunreuther and Michel-Kerjan, 2009). Furthermore, two-thirds of them occurred since 2001. This series of unprecedented catastrophes raises questions: Why is this happening? Will the coming years be even worse? If so, how are individuals—residents and policy makers alike—likely to behave in this new environment? And, finally, what might be done to reduce future losses?

The goal of this chapter is to explore these questions. Our central thesis is that escalating losses from natural hazards are the result of a dynamic interplay between two central forces: one economic, the other behavioral. The economic drivers of catastrophic losses are the increasing amounts of material and numbers of human assets that have been placed in harm's way without adequate compensating investments in mitigation—most notably in coastal areas adjacent to the world's oceans. These location decisions, in turn, arise from a tendency among residents and policy makers to underattend to low-probability, high-consequence risks. While decision makers are quick to see the potential short-term gains that can be obtained from investing in development in such areas, they display less skill at comprehending the long-term risks of such development or of seeing the benefits of long-term investments in protection. Moreover, the rarity of natural disasters in a given location provides few opportunities to correct such mistaken beliefs—until, of course, catastrophes occur. The result is an accelerating spiral of risk taking, in which the rate of economic development in high-risk areas increasingly outpaces technological gains in how to protect those assets, as well as the willingness of residents and policy makers to invest in these technologies.

We divide our discussion of the basis—and possible solutions—to the catastrophic risk problem into three phases. We will first set the stage by reviewing evidence for the most immediate driver of catastrophes, the increasing global mismatch that exists between

assets at risk and investments in risk-reducing measures. We will then explore the underlying behavioral drivers of this mismatch, noting that failures to invest in mitigation are often due to fundamental biases in how we make decisions under uncertainty and plan for the future—biases that can have disastrous consequences when applied to problems involving infrequent natural hazards. We conclude with a discussion of how knowledge of these human tendencies might suggest decision architectures that could help individuals and societies better manage the risk of future catastrophic losses. We focus on the specific example of how myopia biases can be overcome by offering residents and businesses long-term insurance policies coupled with long-term home-improvement loans to induce individuals to invest in cost-effective mitigation measures. We show that this proposal, coupled with well-enforced building codes, significantly improves both individual and social welfare.

The Investment-Mitigation Gap

For evidence of why losses from natural disasters have increased so rapidly in recent years one needs to look no further than the state of Florida. The 1,200 miles of coastline that make it an attractive destination for tourists and retirees also make it vulnerable to impacts by hurricanes from the Atlantic, Gulf of Mexico, and Caribbean. While the omnipresent threat of such hurricanes has long been a part of life in the state, its economic impact was historically limited by the sparseness of the population. As late as 1950 the state was only the twentieth largest in the United States, with a population of 2.8 million. But the years since then have witnessed a migration boom, with the state now being the country's fourth largest, with a population of over 19 million in 2011 (a 600% increase since 1950). The consequence is clear: storms that were previously sources of inconvenience are now potential sources of catastrophe. It has been conjectured, for example, that if the same strong hurricane that hit Miami in 1926 were to hit the same area today, it would induce economic losses that would dwarf those the country recently saw from Hurricane Katrina (Pielke et. al., 2008). But this increased exposure is hardly unique to Florida. As of December 2007, Florida and New York each had nearly $2.5 trillion of insured value located directly on the coast. The coastal insured value for the top ten states combined (ranked by that variable) accounts for more than $8.3 trillion (Kunreuther and Michel-Kerjan, 2009). Such huge concentrations of insured value in highly exposed areas almost guarantees that any major storm that hits these regions will inflict billions, if not hundreds of billions, of dollars of economic losses, unless the residential construction and infrastructures are properly protected by effective mitigation measures.

How well protected are properties in hazard-prone areas? The empirical evidence is disturbing. A 1974 survey of more than a thousand California homeowners in earthquake-prone areas, for example, revealed that only 12% of the respondents had adopted any protective measures (Kunreuther et al., 1978). Fifteen years later, there was little change despite the increased public awareness of the earthquake hazard. In a 1989 survey of 3,500 homeowners in four California counties at risk from earthquakes, only 5%–9% of the respondents in these areas reported adopting any loss-reduction measures (Palm et al., 1990). Burby et al. (1988) and Laska (1991) have found a similar reluctance by residents in flood-prone areas to invest in mitigation measures.

Likewise, even after hurricanes caused extensive damage to large parts of the U.S. Atlantic and Gulf coastlines during the 2004 and 2005 hurricane seasons, a large number of residents had still not invested in relatively inexpensive loss-reduction measures with respect to their property, nor had they undertaken emergency preparedness measures. A survey of 1,100 adults living along the Atlantic and Gulf Coasts undertaken in May 2006 revealed that 83% of the responders had taken no steps to fortify their homes, 68% had no hurricane survival kit, and 60% had no family disaster plan (Goodnough, 2006).

The lack of interest in mitigation measures, even after the most devastating hurricane in the history of the country, is very much puzzling because we know that risk-reduction measures are effective. An analysis of the reduction in damage from future hurricanes in four states (Florida, New York, South Carolina, and Texas) if current building codes were to be applied to all residential property in harm's way is revealing. The reductions range from 61% in Florida for a 100-year return-period loss to 31% in New York for a 500-year return-period loss. In Florida alone, mitigation would reduce losses by $51 billion for a 100-year event and $83 billion for a 500-year event (Kunreuther and Michel-Kerjan, 2009).

What makes matters worse is that this failure to prepare for future disasters has consequences that go beyond the losses suffered by the owners of unprotected structures. When homeowners, private businesses, and the public sector do not adopt cost-effective loss-reduction measures, large sections of coastline are left highly vulnerable to catastrophic losses that can have significant economic spillover effects. These broader losses, in turn, often lead public

sector agencies to provide disaster relief to victims and subsidies to the affected victims even if the government claimed it had no intention of doing so prior to the event. This combination of underinvestment in protection coupled with the general taxpayer financing losses after the fact has been termed the *natural disaster syndrome* (Kunreuther, 1996).

Why We Underprepare: The Psychology of Hazard Prevention

Why do individuals and communities seem so reluctant to invest in mitigation when the long-term benefits are significant? To explore this issue it is useful to begin by reviewing how mitigation decisions *should* ideally be made by a homeowner who makes choices by maximizing expected utility. With this as a benchmark, we will then explore how different psychological tendencies and simplified decision rules foster decisions that depart from economic rationality; we will call those *decision biases.*

Consider the Lowlands, a hypothetical family whose New Orleans home was destroyed by Hurricane Katrina. They have decided to rebuild their property in the same location but are unsure whether they want to invest in a flood-reduction measure (e.g., by elevating their home, sealing the foundation of the structure, and waterproofing the walls).[1] If the flood-proofing measure costs $20,000, should they make the investment?

On the surface, the problem would seem a natural candidate for utilizing expected utility theory. The Lowlands could simplify their decision rule by determining where they should invest in mitigation if they were neutral with respect to risk. If the long-term expected benefits of protection, discounted appropriately to reflect the time value of money, exceeded the up-front costs of the measure, then they should undertake this action. The expected utility model implies that the Lowlands would be even more interested in investing in mitigation if they were averse to the risk of large losses from future disasters.

If the family were to attempt such an analysis, they would quickly realize that they lack most of the critical information needed to make the relevant comparison of costs and benefits. For example, the future economic benefit of mitigation conditional on a flood is highly uncertain. It depends not only on the quality of implementation (which is unobservable) but also on future social and economic factors over which the Lowlands have little control, such as whether neighbors make similar investments that are likely to impact the value of their property, or whether federal disaster relief will be made available following a disaster. The decision is further complicated by the *timing* of the choice; the optimal mitigation policy might be to postpone the investment until the above ambiguities are resolved.

In the absence of analytic guidance, how will the Lowlands make the decision? Central to this chapter is the hypothesis that individuals often utilize informal heuristics that have proven useful for guiding day-to-day decisions in more familiar contexts but that are likely to be unsuccessful when applied to the kind of low-probability, high-stakes decisions they are now facing in a catastrophic environment. In the subsections below, we will review the range of informal mechanisms that are used to make mitigation decisions and discuss how those mechanisms might explain the widespread lack of investment illustrated above.

Budgeting Heuristics

The simplest explanation as to why individuals fail to mitigate in the face of transparent risks is affordability. If the Lowland family focuses on the up-front cost of flood-proofing their house and they have limited disposable income after purchasing necessities, there would be little point in undertaking a comparison of expected benefits and costs regardless of its recommendation. Residents in hazard-prone areas have used this argument explicitly as to why they have limited interest in buying insurance voluntarily. In focus-group interviews to determine the factors influencing decisions about whether to buy flood or earthquake coverage, one uninsured worker responded to the question, How does one decide how much to pay for insurance? by responding as follows:

> A blue-collar worker doesn't just run up there with $200 [the insurance premium] and buy a policy. The world knows that 90 percent of us live from payday to payday. . . . He can't come up with that much cash all of a sudden and turn around and meet all his other obligations." (Kunreuther et al., 1978, p. 113)

A similar argument is likely to be made by individuals when it comes to investing in protective measures such as elevating one's house. In fact, such a budget constraint may extend to higher-income individuals if they set up separate mental accounts for different expenditures (Thaler, 1999). Under such a heuristic, a homeowner who is uncertain about the cost-effectiveness of mitigation might simply compare the price of the measure to what is typically paid for comparable home improvements. Hence, the $20,000 investment may be seen as affordable by those who frame it as a large improvement similar to installing a

new roof, but as unaffordable to those who frame it as a repair similar to fixing a leaky faucet.

Making mitigation decisions in this manner does not conform to the guidelines implied by expected utility theory or cost-benefit analysis, but there is evidence from controlled laboratory experiments that it may not be uncommon. For example, in a study in which individuals were asked why they were willing to pay only a fixed amount for a dead-bolt lock when the lease for the apartment was extended from one to five years, one respondent said, "$20 is all the dollars I have in the short-run to spend on a lock. If I had more, I would spend more—maybe up to $50" (Kunreuther, Onculer, and Slovic, 1998, p. 284). Similarly, we suspect that some residents in coastal zones are discouraged from buying and installing storm shutters because the cost exceeds that of the window itself—a logical benchmark expenditure.

Biases in Temporal Planning

While individuals' decisions about mitigation are undoubtedly constrained by considerations of affordability, trade-offs between costs and benefits invariably arise at some level. Are people skilled at making these comparisons? The empirical evidence on how individuals make intertemporal judgments is not encouraging. Although decisions often follow the directional advice of normative theory (such as by valuing temporally distant events less than immediate ones), they frequently depart from those prescribed by rational theories of intertemporal choice. Moreover, they depart in a way that collectively discourages farsighted investments in mitigation.

To understand this, consider the investment problem faced by the Lowlands. For simplicity, suppose that the family knows that they will be living in their new home for T years, that each year there is a probability p_t of a Katrina-like flood in year t, and that should such an event occur, the mitigation measures will reduce losses by an amount B. In this case, the decision to mitigate could be made by observing whether the disutility associated with the upfront cost (C) of mitigation is less than the positive utility associated with the discounted stream of benefits; that is, if

$$u(C) < \sum_{t=1}^{T} p_t u(B)\beta^t \qquad (23.1)$$

where β is the consumer's discount rate and $u(x)$ is the consumer's utility associated with the benefit (B) or cost (C).

While simple in its structure, implicit in equation 23.1 is a series of rather strong assumptions about how the Lowlands will value costs and benefits over time; specifically:

1. all future benefits are evaluated vis-a-vis a constant rate of discounting;
2. individuals can estimate future probabilities of flooding in year t accurately; and
3. the utility function is time-invariant.

There is ample evidence that violations of these assumptions will be common. In particular, homeowners are likely to overweight short-term cash expenditures, have distorted beliefs about probabilities, and value common outcomes differently over time. The implications of these tendencies in the context of mitigation decisions are now reviewed in turn.

UNDERWEIGHTING THE FUTURE

A fundamental feature of human cognition is that we are influenced more by cues that are concrete and immediate than those that are abstract and delayed. To some extent, of course, rational intertemporal choice theory prescribes that we *should* give less weight to distant future outcomes, and this prescription is captured by the constant discount rate β in equation 23.1. There is extensive experimental evidence, however, that human temporal discounting tends to be *hyperbolic*, where temporally distant events are disproportionately discounted relative to immediate ones. As an example, people are willing to pay more to have the timing of the receipt of a cash prize accelerated from tomorrow to today than from two days from now to tomorrow (in both cases a one-day difference) (Loewenstein and Prelec, 1992). The implication of hyperbolic discounting for mitigation decisions is that residents are asked to invest a tangible fixed sum now to achieve a benefit later that they instinctively undervalue—and one that they, paradoxically, hope never to see at all.

The effect of placing too much weight on immediate considerations is that the up-front costs of mitigation will loom disproportionately large relative to the delayed expected benefits in losses over time. A homeowner might recognize the need for mitigation and see it as a worthwhile investment when it is framed as something to be undertaken a few years from now when both the up-front costs and the delayed benefits are equally discounted. However, when the time arrives to actually make the investment, a homeowner subject to hyperbolic discounting might well get cold feet.

This tendency to shy away from undertaking investments that abstractly seem worthwhile is exacerbated if individuals have the ability to *postpone* investments—something that would almost always be the case with respect to mitigation. A case in point is the relative lack of preparedness demonstrated by

the city of New Orleans and the Federal Emergency Management Agency (FEMA) in advance of Hurricane Katrina in 2005. In this case, the consequences of failing to invest in mitigation—such as developing a workable evacuation plan—could not have been more salient or more temporally proximate. Just two months prior to the storm, the city engaged in a full-scale simulation that graphically demonstrated what would happen should a hurricane of Katrina's strength hit the city, and the city was moving into the heart of an active hurricane season (Brinkley, 2006). Yet, little was done to remedy known flaws in their preparedness plans.

What explains the inaction? The explanation, we suggest, is simple: while emergency planners and the New Orleans mayor's office were fully aware of the risks the city faced and understood the need for investments in preparedness, there was inherent ambiguity about just *what* these investments should be and *when* they should be undertaken. Faced with this uncertainty, planners did what decision makers tend to do when faced with a complex discretionary choice: they opted to defer it to the future, in the (usually false) hope that the correct choices would become clearer or more resources would then be available or both (Tversky and Shafir, 1992).

To see this effect more formally, imagine the Lowlands view the future benefits of mitigation not in terms of a constant discounting schedule, but instead by the hyperbolic discounting function

$$f(t) = \begin{cases} 1/k \ \textit{for}\ t = 0 \\ \beta^t \ \textit{for}\ t > 0 \end{cases} \qquad (23.2)$$

where $0 < k < 1$ is a constant that reflects the degree to which immediate costs and benefits are given disproportionately more weight than delayed ones (Laibson, 1997; Meyer, Zhao, and Han, 2007). Equation 23.2 has an intriguing implication. Suppose that it is January ($t = 0$) and the Lowlands are considering whether it is worthwhile to invest in a mitigation project that would start the next June ($t = 1$). As long as costs remain temporally distant, the value of the project will be assessed via the rational intertemporal discounting model in equation 23.1, that is, the expected net value of the mitigation project, next January is

$$V(I \mid \textit{January}) = \left[\sum_{t=1}^{T} p_t k \beta^t u(B)\right] - \beta u(C) \qquad (23.3)$$

Suppose the Lowlands conclude that the project is minimally worthwhile a year from now, that is, $V(I|\text{January}) = \varepsilon$, where ε is a small positive valuation. Hyperbolic discounting carries a curious implication for how the Lowlands will value the project come July, when the prospect of the expenditure C is immediate. In June, the project will look decidedly less attractive, since its value will now be

$$V(I \mid \textit{June}) = \left[\sum_{t=1}^{T} p_t u(B) k \beta^t\right] - u(C)/k \qquad (23.4)$$

Hence, if $(1/k - \beta)C > \varepsilon$, it will no longer seem worthwhile to invest. So will the Lowlands abandon their interest in mitigation? Paradoxically, we suggest no; if the builder gives them the option to restart the project the *following* January, it will once again seem worthwhile, since its valuation would be given by the standard model in equation 23.3. Hence, the Lowlands would be trapped in an endless cycle of procrastination; when viewed from a temporal distance, the investment will always seem worthwhile, but when it comes time to undertaking the work, the prospect of a slight delay always seems more attractive.

We should add that other, less formal psychological mechanisms could also produce perpetual postponements of investments in mitigation. The most salient is the observed tendency for individuals to defer ambiguous choices; the less certain one is about a correct course of positive action, the more likely one is to choose inaction (Tversky and Shafir, 1992). This ambiguity would seem particularly acute in the context of mitigation decisions, where the question of whether it is optimal to mitigate is often unknowable for a single household and there is infinite flexibility as to *when* one can undertake the investment. Finally, when viewed locally, the risk of a short delay in the start of mitigation is typically negligible. While seismologists are reasonably certain that there will be a major quake along the San Andreas Fault in southern California at some point over the next century, odds are strongly against it happening tomorrow. As such, residents who postpone the decision from day to day will rarely be punished for their inaction.

We should emphasize that the concept of hyperbolic discounting discussed above is distinct from that of *planning myopia*, the tendency to consider consequences over too short a finite time horizon. For example, if the Lowlands' beliefs about the length of time they would live in their home were biased downward, they would underestimate the benefits of mitigation by using equation 23.1. While we are not aware of work that has examined whether there are systematic tendencies to misjudge homeownership tenure, the fact that the vast majority (72%) of U.S. homeowners prefer thirty-year fixed (as opposed to adjustable) mortgages has been taken by some economists as evidence that homeowners, if anything, *over*estimate the length of time they will likely live in their homes (Campbell, 2006). It is thus paradoxical, then, that homeowners would display acute concern for minimizing long-term risk when securing mortgages,

but display little comparable concern when making decisions about investing in mitigating potential damage to their home.

UNDERESTIMATION OF RISK

Another factor that could suppress investments in mitigation is underestimation of the likelihood of a hazard—formally, underestimation of p_t in equation 23.1. Although underestimation of risk is perhaps the simplest explanation as to why people fail to mitigate, the empirical evidence in the domain of natural hazards is far more complex.

On the one hand, we do know that decisions about mitigation are rarely based on formal beliefs about probabilities. Magat, Viscusi, and Huber (1987) and Camerer and Kunreuther (1989), for example, provided considerable empirical evidence that individuals do not seek out information on probabilities in making their decisions. Huber, Wider, and Huber (1997) showed that only 22% of subjects sought out probability information when evaluating risk managerial decisions. When consumers are asked to justify their decisions on purchasing warranties for products that may need repair, they rarely use probability as a rationale for purchasing this protection (Hogarth and Kunreuther, 1995).

Even though individuals do not find statistical probability to be a useful construct in making risky decisions, they are able to provide estimates of their subjective beliefs about relative risk. But these beliefs are not well calibrated. When directly asked to express an opinion about the odds of being personally affected by different hazards, people consistently respond with numbers that, perhaps surprisingly, are far *too high* relative to actuarial base rates. For example, in a study of risk perception, Lerner et al. (2003) found that when people were asked to provide an estimate of the probability that they will be the victim of a violent crime over the coming year, the mean estimate was 43%—an estimate that was far too high compared to actuarial base rates and comparable to that which they expressed when asked to estimate the odds of getting the flu (47%). If these estimates actually reflected heightened fears about being exposed to hazards, it would strongly argue against the idea that people fail to mitigate simply because they assume that they will be immune. But these results may be speaking more to individuals' lack of familiarity with statistical constructs than to real evidence that people are pessimistic.

On the other hand, there is also evidence that people tend to ignore risks whose subjective odds are seen as falling below some threshold. In a laboratory experiment on purchasing insurance, many individuals bid zero for coverage, apparently viewing the probability of a loss as sufficiently small that they were not interested in protecting themselves against it (McClelland, Schulze, and Coursey, 1993). Similarly, many homeowners residing in communities that are potential sites for nuclear waste facilities have a tendency to dismiss the risk as negligible (Oberholzer-Gee, 1998). Prior to the Bhopal chemical accident in 1984, firms in the industry had estimated the chances of such an accident as sufficiently low that it was not on their radar screen (Bowman and Kunreuther, 1988). Similarly, even experts in risk disregard some hazards. For instance, even after the first terrorist attack against the World Trade Center in 1993, terrorism risk continued to be included as an unnamed peril in most commercial insurance policies in the United States, so insurers were liable for losses from a terrorist attack without their ever receiving a penny for this coverage. (Kunreuther and Pauly, 2005). Because insurers had not integrated the threat into their portfolio management, the September 11, 2001, attacks obligated them to pay over $35 billion in claims.

Levees or other flood control projects are likely to give residents a false sense of security with respect to suffering damage from floods or hurricanes. In fact, Gilbert White (1975) pointed out many years ago that when these projects are constructed, there is increased development in these "protected" areas. Should a catastrophic disaster occur so that residents of the area are flooded, the damage is likely to be considerably greater than before the flood-control project was initiated. This behavior and its resulting consequences have been termed the *levee effect*. Evidence along these lines has more recently been offered by Burby (2006), who argued that actions taken by the federal government, such as building levees, make residents feel safe when, in fact, they are still targets for catastrophes should the levee be breached or overtopped.

AFFECTIVE FORECASTING ERRORS

A final assumption of normative theories of intertemporal choice that is worth scrutinizing is the assumption that utility functions are temporally invariant. In our example, the Lowlands would be assumed to value benefits from mitigation realized in the distant future in the same way that they would be valued if realized now. How likely is this assumption to be empirically valid? There are extensive bodies of work showing that individuals tend to be both poor forecasters of future affective states (e.g., Wilson and Gilbert, 2003) and focus on different features of alternatives when they are viewed in the distant future versus today (e.g., Trope and Liberman, 2003).

Probably the most problematic of these biases for mitigation decisions is the tendency for

affective forecasts to be subject to what Loewenstein, O'Donoghue, and Rabin (2003) term the *projection bias*—a tendency to anchor beliefs about how we will feel in the future on what is being felt in the present. Because mitigation decisions are ideally made in docile times, long before (rather than just after) a disaster occurs, the projection bias predicts a tendency for decision makers to both underestimate the likelihood of future hazards and the feelings of trauma that such events can induce—a bias that would, in turn, lead to undervaluation of investments in protection. After Hurricane Katrina, a common theme heard from survivors trapped in the floods was, "Had I known it would be this bad, I would have left." The reality, of course, was that they were *told* that it would be that bad; the storm was preceded by warnings of the most dire sort, that Katrina was "the big one" that New Orleans' residents had been warned to fear for years (Brinkley, 2006). But it is one thing to imagine being in a large-scale flood, quite another to actually be in one. Judgments of the severity of the predicted experience were unavoidably biased downward by the relative tranquility of life before the storm.

We might add that while the likely dominant effect of affective forecasting errors is to undervalue future protection, it is possible that protection could be *over*valued if prior valuations are subject to a different bias that has also been observed in intuitive forecasts, that of *duration neglect*, the tendency to overestimate the length of time it takes to recover from negative life events, such as being fired from a job (Wilson and Gilbert, 2003). Under such a bias, a homeowner might well underestimate the initial impact of damage suffered from a hazard due to the projection biases, but still *over*invest in protection out of a belief that it will take an excessively long time to physically and emotionally recover from that damage. As an example, in the days immediately following Katrina, there were dire warnings that it would likely be months before flooded sections of the city could be drained and that the city would never be able to recover—predictions that later proved too pessimistic.[2]

Finally, the tendency to value costs and benefits differently depending on temporal perspective is another mechanism that could result in procrastination. Trope and Liberman (2003) offer a wide array of evidence showing that when making choices for the distant future, we tend to focus on the abstract benefits of options, whereas when making immediate choices, we tend to focus on concrete costs. Hence, similar to the predictions made by hyperbolic discounting, it would not be uncommon to hear politicians pledge their deep commitment to building safer societies at election time (when costs loom small relative to abstract benefits) but then back away from this pledge when the time comes to actually make the investment—when it is the concrete costs that loom larger.

Learning Failures

The above discussion makes a clear argument that if individuals make mitigation decisions by performing intuitive comparisons of up-front costs with long-term benefits, they will likely underinvest by virtue of focusing too much on up-front costs, undervaluing long-term benefits or underestimating the likelihood that the disaster will happen to them or both. But this begs a conjecture: while an individual (or institution) making a one-time mitigation decision might well err by underinvesting, such errors would likely be transient. Once the consequences of undermitigation are observed, intuition suggests that there would be a natural tendency to correct the biases that led to the initial error. Indeed, there is some evidence that mitigation errors are naturally correcting; the early Mayans learned (no doubt by experience) that it was safer to build cities inland than on the hurricane-prone coasts of the Yucatan. The loss of 6,000 lives in Galveston in 1900 taught the city that it needed a seawall to protect against future storms, and it took the disaster of Katrina for New Orleans to finally put in place a comprehensive evacuation plan (Brinkley, 2006).

The problem, however, is not that we do not learn, but rather that we do not seem to learn *enough* from the experiences of disaster. As an illustration, when Hurricane Wilma hit south Florida in October 2005, just a few weeks after Hurricane Katrina, thousands of residents failed to take such simple preparatory measures as securing bottled water and filling their cars up with gas—oversights that greatly added to the challenges of recovery. What was surprising about this lack of preparation was that the region had ample warning of the storm's approach (the impact was forecast up to four days in advance), and it came at the end of the most destructive hurricane season on record, one where the news media were constantly filled with graphic illustrations of the destructive power of such storms (such as the flooding in New Orleans). Other familiar examples exist as well—such as the tendencies to resettle in flood plains after major floods and to become increasingly lax in earthquake preparedness as the time since the last quake lengthens.

What explains the seeming lack of learning by residents? The reason, we suggest, is that we instinctively learn to protect ourselves against hazards by relying on the same trial-and-error heuristics that have proven successful in other walks of life: heuristics that encourage us to repeat those behaviors that yield positive rewards and avoid those behaviors that yield negative outcomes. But while reinforcement learning

is a highly efficient way to learn to perform such repeated tasks as walking, speaking, and playing tennis, it is particularly ill-suited for learning how best to respond to low-probability, high-consequence hazards. The reason is simple: most protective behaviors will be negatively reinforced far more often than they will be positively reinforced.

As an example, when Hurricane Wilma approached south Florida in 2005, the vivid news depictions of suffering Katrina survivors were counterbalanced by a different and more salient source of information: residents' recollections of the seven false alarms that the area had received during the previous two years. For many, the hurricane warnings posted for Wilma triggered memories of rushing to secure supplies of water and gas before a storm, only later to find out that their efforts were unnecessary. For everyone else, it was memories of how gambling paid off; their decisions *not* to prepare for all the *previous* storms had turned out to be the right ones (in hindsight).

A more recent example of the potential dysfunctionalities of learning from experience arose in a different context, too: the 2007 shootings at Virginia Tech University. In a report that was highly critical of the university's slow response to the incident, an investigatory committee suggested that the failure to react quickly was partially due to the bad publicity that the university administration received the previous year when they were accused of *over*reacting to the threat of an escaped convict who was involved in a shooting near campus. The threat proved a false alarm, but there was nevertheless damage: administrators became wary of moving too quickly in response to the next threat.

This same tendency to ignore warning information because of its perceived unreliability may also be inflamed in the context of natural hazards by the tendency for news sources to overhype threats—such as warnings of an impending blizzard, ice storm, or hurricane. While occasionally the events measure up to the billing as advertised on the 11 p.m. news, more often they will not—causing viewers to discount future warnings of dire threats.

A second major impediment to learning is the inherent ambiguity of feedback about what constitutes optimal mitigation. In the course of disasters, one can rarely observe the counterfactuals critical to learning: what damage would have occurred had certain mitigation steps been taken (or not taken). As noted by Meyer (2006), one consequence of this feedback property is that it supports the persistence of superstitious beliefs about mitigation strategies. A good example is the old adage that one should open windows in advance of a tornado so as to equalize pressure. It took structural engineers years to discover that open windows were more likely to be the *cause* of building failures than the cure (entering wind exerts upward pressure on roofs); yet the myth is still widely held. The reason, of course, is that it is impossible to infer from observing a destroyed house whether it would still be standing had the windows been closed—or indeed whether they were open or closed to begin with.

We should emphasize that it is not the rarity of hazardous events per se that limits learning. While individuals may encounter a major earthquake or hurricane only once in their lives (or, more likely, never), there are ample opportunities to learn by observing the experiences of others. Indeed, the plethora of books describing great disasters of the past (from Noah's flood onward) and the intense new attention that is given to disasters suggests that we have deeply ingrained instincts (however morbid) for trying to learn from others' misfortunes. There is suggestive evidence, however, that people often learn much less from vicarious feedback than one might hope. In one example, a laboratory study designed to measure peoples' abilities to learn optimal levels of investment to protect against hurricanes, Meyer (2006) found that decisions to increase investment were driven almost exclusively by whether the decision maker personally suffered losses in the previous period; in contrast, losses suffered by others did not have such a triggering effect.

Finally, we should note that if the government comes to the rescue with liberal disaster assistance, then one may conclude from media reports or personal experience that it may not be necessary to invest in costly protective measures. In fact, those who have taken steps to protect themselves financially against losses may conclude that they would be better off not having purchased coverage. A graphic example comes from the Alaska earthquake of 1964, when the federal government provided low-interest loans to aid the recovery and retire debts from existing mortgages for those who were uninsured. It was not uncommon to hear the few homeowners who did purchase earthquake insurance bemoan their decision because they discovered they would have been better off financially had they not purchased this coverage. (Dacy and Kunreuther, 1968).

Social Norms and Interdependencies

Let us return again to the dilemma faced by the Lowland family, who have now narrowed their mitigation option to elevating their house on piles so as to reduce flood losses from a future hurricane. If none of their neighbors have taken this step, their house would look like an oddity in a sea of homes at ground level. Should the Lowlands choose to move, they would be

concerned that the resale value of their home would be lower because the house was different from all the others. Likewise, the effectiveness of mitigation might itself depend on the number of other residents who elect to elevate their homes. If the Lowlands decide to elevate their home but their neighbors do not take this action, then one of these nonelevated homes could be washed away during a flood or hurricane and cause damage to the Lowlands' home which would otherwise have been spared. Note that these—very real—considerations would not be easily captured in a traditional expected utility analysis of their problem (such as equation 23.1), which assumes a decision is made in social isolation. In contrast, mitigation decisions often take the form of coordination games, where the value of mitigation depends on whether neighbors choose to mitigate.

Decisions made by neighbors also carry information value—or at least are likely to be perceived as such. As in an information cascade (Sunstein, 2006), if a large number of neighbors have already decided to put their houses up on piles, the Lowlands might plausibly conclude that the investment must be cost effective. Of course, such inferences could be wildly mistaken if their neighbors' decisions were also based on imitation; much like a fad, one might observe communities collectively adopting mitigation measures that have little actuarial or engineering basis.

To illustrate such effects, we recently conducted a laboratory study of social-network effects in earthquake mitigation. In the study, participants were told that they would be living in an area prone to periodic earthquakes and that they could purchase structural improvements in their homes that potentially mitigated the effects of quakes should one arise. The task was to make these decisions as efficiently as possible in the following sense: at the end of the simulation they would be paid an amount that was tied to the difference between their home value and interest earned minus the cost of mitigation plus damage repairs. Throughout the simulation they could observe the investment decisions being made by others in their virtual community, as well as the damage they suffered from quakes. The key source of uncertainty in the simulation was whether the mitigation was cost effective or not; half of the participants were placed in a world where mitigation was not cost effective (hence the optimal investment was 0%), and the other half were placed in a world where it was long-term effective (hence the optimal investment was 100%). Our interest was in observing whether communities could discover the optimal level of mitigation over repeated plays of the game.

The basic result was that they could not determine how much to invest in mitigation. Consistent with the findings on learning discussed above, there was little evidence of either community naturally discovering the optimal level of mitigation (the investment level in both worlds averaged 40%). There was, however, a social norm effect: the major driver of individual decisions about how much to invest was the average level of investment made by neighbors.

Would their learning have been enhanced had the communities been populated with a few opinion leaders who had knowledge of mitigation's true effectiveness? To investigate this, we ran a new set of studies where, prior to the simulation, one player in each community was privately informed of the true effectiveness of mitigation. Other players knew that one among them had this information, but that person's identity was not revealed but could likely be inferred by observing players' investment behavior. For example, a player who is told that investments are ineffective would, presumably, invest 0% from the start. Did this "knowledge seeding" help communities learn? It did, but—quite surprisingly—only in the case where investments were ineffective. In these communities, players seemed to immediately recognize the informed player (who was not investing), and after two rounds of the game almost all investments in mitigation had vanished, as it should have.

In contrast, in communities where mitigation was effective, rather than investments increasing over time, they *decreased*. For many of the reasons described earlier in this section, players who were told that mitigation was effective did not play the optimal strategy of investing 100% at the start—they procrastinated. The other players, seeing no one with a high level of investment, then mistakenly concluded that they must be in a world where mitigation was ineffective. Hence they invested only a small amount themselves. Then, bizarrely, the informed players—who should have been opinion leaders—became followers, reducing their own investments. After multiple plays of the game, few players were making any investments at all, even though it was optimal for them to do so in the long run.

Of course, one might hope that in real-world settings, opinion leadership and tipping strategies might be more effective, and evidence along these lines has been presented by Schelling (1978) and popularized by Gladwell (2000). Heal and Kunreuther (2005) provided a game theoretic treatment of the impact of interdependency on the decision to invest in protective measures and suggested ways to coordinate actions of those at risk, ranging from subsidization or taxation to induce tipping or cascading to rules and regulations, such as well-enforced building codes.

The Samaritan's Dilemma

As alluded to above, another of the arguments advanced as to why individuals do not adopt protective

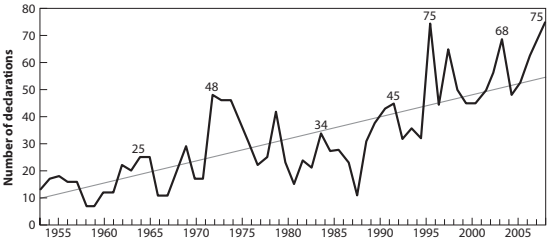
measures prior to a disaster is that they assume that liberal aid from the government will be forthcoming should they suffer losses from an earthquake, hurricane, or flood. Federal disaster assistance may create a type of Samaritan's dilemma: providing assistance *ex post* (after hardship) reduces parties' incentives to manage risk *ex ante* (before hardship occurs) (Buchanan, 1975). If the Lowland family expects to receive government relief after a loss, it will have less economic incentive to invest in mitigation measures and purchase insurance prior to a hurricane. The increased loss due to the lack of protection by residents in hazard-prone areas amplifies the government's incentive to provide assistance after a disaster to victims.

The empirical evidence on the role of disaster relief suggests, however, that individuals or communities have *not* based their decisions on whether or not to invest in mitigation measures by focusing on the expectation of future disaster relief. Kunreuther et al. (1978) found that most homeowners in earthquake- and hurricane-prone areas did not expect to receive aid from the federal government following a disaster. Burby (1991) found that local governments that received disaster relief undertook more efforts to reduce losses from future disasters than those that did not. This behavior seems counterintuitive and the reasons for it are not fully understood. It will be interesting to see whether Hurricane Katrina changes this view, given the billions of dollars in disaster relief that have been sent to victims and affected states. In the same vein, the historical federal bailout of some of the largest financial institutions in 2008 and 2009 is likely to create strong moral hazard in the future.

Whether or not individuals incorporate an expectation of disaster assistance in their predisaster planning process, a driving force with respect to the actual provision of government relief is the occurrence of disasters where the losses are large (Moss, 2002). Under the current system, the governor(s) of the state(s) can request that the president declare a "major disaster" and offer special assistance if the damage is severe enough. Although the president does not determine the amount of aid (the House and Senate do), he is responsible for a crucial step in the process. This obviously raises the questions, What are the key drivers of such a decision? Are some states more likely to benefit from this situation than others, and if so, when does this occur?

The Politician's Dilemma

Federal relief is not immune from behavioral bias—that observation is consistent with recent research that has shown that election years are a very active time for disaster assistance (all other things being equal). Figure 23.1 depicts the evolution of the number of presidential declarations over the period 1953–2008. Overall, the number of these declarations has dramatically increased during that time: there were 162 during 1955–1965, 282 between 1966 and 1975, 319 over the period 1986–1995, and 545 from 1996 to 2005 (Michel-Kerjan, 2008b). It is interesting to note that many of the peak years described in figure 23.1 correspond to presidential election years.



23.1. Presidential disaster declarations per year (peak values on the graph correspond to some presidential election years).
(Redrawn from Michel-Kerjan, 2008b)

Four salient examples are the Alaska earthquake in 1964, Tropical Storm Agnes in June 1972, Hurricane Andrew in September 1992, and the four hurricanes in 2004. For example, following the Alaska earthquake, when relatively few homes and businesses had earthquake-resistant measures and insurance protection, the U.S. Small Business Administration provided 1% loans for rebuilding structures and refinancing mortgages to those who required funds through its disaster loan program. As pointed out above, the uninsured victims in Alaska were financially better off after the earthquake than their insured counterparts (Dacy and Kunreuther, 1968). More recently, it has also been shown that a battleground state with twenty electoral votes has received more than twice as many presidential disaster declarations than a state with only three electoral votes (Reeves, 2004, 2005).

In the case of Hurricane Katrina, Governor Kathleen Blanco declared a state of emergency on August 26, 2005, and requested disaster relief funds from the federal government on August 28. President Bush declared a state of emergency on the twenty-eighth, an action that freed federal government funds and put emergency response activities, debris removal, and individual assistance and housing programs under federal control (Congressional Research Service, 2005). Under an emergency declaration, federal funds were capped at $5 million. On August 29, in response to Governor Blanco's request, the president declared a "major disaster," allotting more federal funds to aid in rescue and recovery. By September 8, Congress had approved $52 billion in aid to the victims of Hurricane Katrina. As of August 2007, the total federal

relief allocated by Congress for the reconstruction of the areas devastated by the 2005 hurricane season was nearly $125 billion.

The fact that politicians can benefit from their generous actions following a disaster raises basic questions as to the capacity of elected representatives at the local, state, and federal levels to induce people to adopt protection measures before the next disaster. The difficulty in enforcing these mitigation measures has been characterized as the *politician's dilemma* (Michel-Kerjan, 2008a).

Imagine an elected representative at the city or state level. Should she push for people and firms in her district to invest in cost-effective mitigation measures to prevent or limit the occurrence of a disaster? From a long-term perspective, the answer should be yes. But given short-term reelection considerations (another form of myopia), the representative is likely to vote for measures that allocate taxpayers' money elsewhere that yield more political capital. It is another example where little consideration is given to supporting mitigation measures prior to a disaster (*ex ante*) because their constituencies are not worried about these events occurring and because there is likely to be a groundswell of support for generous assistance to victims from the public sector after a disaster (*ex post*) to aid their recovery. The one silver lining to this behavior is that following a natural disaster, when residents and the media focus on the magnitude of the losses, politicians are likely to respond by favoring stronger building codes and other loss-reduction measures, but only when there is a consensus among their constituencies that this is a good thing to do.

Strategies for Overcoming Decision Biases

How might we build more resilient communities in areas that are prone to natural hazards? The biases discussed above suggest that the task is not an easy one. Communities face a difficult choice: either find ways to debias decision makers so as to foster voluntary investments in mitigation, or restrict voluntary choice, such as imposing well-enforced building codes and land-use regulations. To date, public officials have turned to regulations as the only effective means of insuring mitigation: if residents are unable to make wise choices about where to live and how much to invest in protection, it is the role of government to impose these choices.

A compelling illustration of this argument was provided by Kydland and Prescott (1977), who, in their Nobel Prize–winning contribution, showed that a policy that allows freedom of choice may be socially optimal in the short run but socially suboptimal from a long-term perspective. As a specific example, the authors noted that unless individuals were initially prohibited from locating in a flood plain, it would be very difficult politically to force these people to leave their homes. Kydland and Prescott argued that these individuals, in making their decisions to locate in flood plains, believed that the Corps of Engineers would subsequently build dams and levees if enough people chose to build homes in flood-prone areas and hence decided to locate there.

Kunreuther and Pauly (2006) extended the Kydland-Prescott argument by introducing behavioral considerations into the picture. They contended that if individuals underestimate the likelihood of a future disaster, it may be important to require homeowners to purchase insurance and have well-enforced rules, such as land-use regulations and building codes, to avoid the large public-sector expenditures following these events. To support this point, they provided empirical evidence that many individuals do not even think about the consequences of a disaster until after a catastrophe occurs and hence do not invest in protective measures in advance of a disaster.

Yet, a policy of widespread government intervention in mitigation decisions carries its own risks. Specifically, the approach can be criticized as one that, paradoxically, might in some cases actually exacerbate rather than reduce the long-term risk of catastrophes. Evidence along these lines has been offered by Burby (2006), who argued that actions taken by the federal government, such as building levees, make residents feel safe when, in fact, they are still targets for catastrophes should the levee be breached or overtopped. Likewise, Steinberg (2000) noted that beach restoration projects that are now widespread in coastal communities carry the same risk: while restoration lowers the potential damage from a single storm event, it also denies residents the visual cues that would inform perceptions of risk. This approach is also one whose viability assumes that the government planners who design and enforce codes are themselves immune from the biases that they are designed to overcome—a presumption that is unlikely to hold in practice. As an example, four decades of hurricanes, from the 1920s through the 1950s, persuaded South Florida to pass one of the country's strictest building codes in 1957, but enforcement of these codes gradually waned during the three quiet decades that followed—a lapse that contributed to the extreme property losses the area suffered during Hurricane Andrew in 1992 (Steinberg, 2000).

In light of this dilemma, a long-term solution to managing catastrophe risks lies in decision architectures that *guide* residents to making more efficient protection decisions in a way that takes into account the behavioral biases noted above.

Long-Term Insurance and Long-Term Mitigation Loans

A major reason why individuals are reluctant to make major investments in mitigation is that they are unlikely to realize personal financial benefits in the short run. A key challenge with respect to encouraging mitigation, therefore, is to move the focal point from the *individual* to the *property*, so that individuals will focus on a longer time horizon when deciding whether to invest in risk-reducing measures. Mortgages can play an important role in this regard since they are typically long-term contracts.

In this spirit we have proposed as a market innovation that one moves from the traditional one-year insurance contracts as we know them today, which encourage myopic thinking, to multiyear insurance contracts with annual premiums coupled with long-term mitigation loans (Jaffee, Kunreuther, and Michel-Kerjan, 2008).

For a long-term insurance (LTI) policy to be feasible (say for 10 or 25 years), insurers would have to be able to charge a rate that reflects their best estimate of the risk over that time period. The uncertainty surrounding these estimates could be reflected in the annual premium being a function of the length of the insurance contract, in much the same way that the interest rate on fixed-rate mortgages varies between 15-, 25-, and 30-year loans. The obvious advantage of an LTI contract from the point of view of policyholders is that it provides them with stability and an assurance that their property is protected for as long as they own it. This has been a major concern in hazard-prone areas, where insurers had cancelled policies before severe disasters occurred.[3] On a much broader scale, a study of flood insurance in Florida revealed that of the 1 million residential National Flood Insurance Program (NFIP) policies in place in Florida in 2000, one-third had been cancelled by 2002 and about two-thirds had been cancelled by 2005 (Michel-Kerjan and Kousky, 2010).

Under the current state-regulated arrangements, where many insurance commissioners have limited insurers' ability to charge premiums that reflect the exposure and the cost of capital necessary to provide insurance in hazard-prone areas, no insurance company would even entertain the possibility of marketing an LTI policy. Insurers would be concerned about the regulator clamping down on them now or in the future regarding what price they could charge, so that LTI insurance would be infeasible from a financial point of view. Given the existing tension between state insurance regulators and the insurance industry, we feel that it would be best politically to introduce LTI by focusing on flood insurance, which is provided nationwide by the federal government under the NFIP.[4]

We propose that the purchase of long-term flood insurance be required for all property owners residing in hazard-prone areas so as to reduce the likelihood of liberal disaster-relief legislation following the next major catastrophe. There is a precedent for such a requirement today in all states where motorists have to show proof of automobile insurance covering bodily injury and property damage liability or financial responsibility in order to register their car. With respect to property insurance, homeowners who have a mortgage are normally required by the bank that finances the loan to purchase coverage against wind damage for the length of the mortgage, and this requirement is normally enforced.

A long-term flood-insurance contract would also provide economic incentives for investing in mitigation where current annual insurance policies are unlikely to do the trick, even if they were risk-based due to the behavioral considerations discussed in the previous section. To highlight this point, consider the following simple example. Suppose the Lowland family could invest $1,500 to floodproof the foundation of its house so as to reduce the water damage by $30,000 from a future flood or hurricane that would occur with an annual probability of 1/100. If the insurance premium reflects the risk, then the annual premium would be reduced by $300 to reflect the lower expected losses to the property. If the house was expected to last for ten or more years, the net present value of the expected benefit of investing in this measure in the form of lower insurance premiums would exceed the up-front cost at an annual discount rate as high as 15%.

Under the current annual flood-insurance contract, many property owners would be reluctant to incur the $1,500 because they would only get $300 back next year. If they used hyperbolic discounting or they were myopic with respect to the time horizon, the expected discounted benefits would likely be less than the $1,500 up-front costs. In addition, budget constraints could discourage them from investing in the mitigation measure. Other considerations could also play a role in the homeowners' decisions not to invest in these measures. They may not know how long they will reside in the area or whether they would be rewarded again with lower premiums next year when their policy is renewed or both.

With a 20-year flood insurance contract, the premium reduction would be viewed as a certainty. In fact, the property owner could take out a $1,500 home-improvement loan tied to the mortgage at an annual interest rate of 10%, resulting in payments of $145 per year. If the annual insurance premium was reduced by $300, the savings to the homeowner each year would be $155.

A bank would have a financial incentive to provide this type of loan. By linking the expenditure in mitigation to the structure rather than to the property owner, the annual payments would be lower, and this would be a selling point to mortgagees. The bank would also feel that it is now better protected against a catastrophic loss to the property, and the insurer knows that its potential loss from a major disaster is reduced. These mitigation loans would constitute a new financial product. Moreover, the general public would now be less likely to have large amounts of their tax dollars going for disaster relief. A win-win-win-win situation for all! (Kunreuther, 2006).

Seals of Approval

A complementary way of encouraging the adoption of cost-effective mitigation measures is to require that banks and other lenders condition their mortgages. Sellers or buyers of new or existing homes would have to obtain a seal of approval from a recognized inspector that the structure meets or exceeds building-code standards. This requirement either could be legislated or imposed by the existing housing government sponsored enterprises (i.e., Fannie Mae, Freddie Mac, and the twelve Federal Home Loan Banks). Existing homeowners might want to seek such a seal of approval as well, if they knew that insurers would provide a premium discount (akin to the discounts that insurers now make available for smoke detectors or burglar alarms) and if home improvement loans for this purpose were generally available.

Evidence from a July 1994 telephone survey of 1,241 residents in six hurricane-prone areas on the Atlantic and Gulf Coasts provides support for some type of seal of approval. Over 90% of the respondents felt that local home builders should be required to follow building codes, and 85% considered it very important that local building departments conduct inspections of new residential construction. We recommend the following procedure. The inspection required to establish a seal of approval must be undertaken by certified contractors. For new properties, the contractor must provide the buyer with this seal of approval. For existing properties, the buyer should pay for the inspection and satisfy the guidelines for a seal of approval. If the house does not satisfy the criteria, then banks and other mortgage lenders should roll into their mortgage loans the cost of such improvements (Kunreuther and Michel-Kerjan, 2006).

Tax Incentives

One way for communities to encourage residents to pursue mitigation measures is to provide them with tax incentives. For example, if a homeowner reduces the losses from a disaster by installing a mitigation measure, then this taxpayer would get a rebate on her state taxes to reflect the lower costs of disaster relief. Alternatively, property taxes could be reduced. But in practice, some communities often create a monetary disincentive to invest in mitigation. A property owner who improves a home by making it safer is likely to have the property reassessed at a higher value and, hence, have to pay higher taxes. California has recognized this problem, and in 1990 voters passed Proposition 127, which exempts seismic rehabilitation improvements to buildings from reassessments that would increase property taxes.

The city of Berkeley in California has taken an additional step to encourage home buyers to retrofit newly purchased homes by instituting a transfer-tax rebate. The city levies a 1.5% tax on property-transfer transactions; up to one-third of this amount can be applied to seismic upgrades during the sale of property. Qualifying upgrades include foundation repairs or replacement, wall bracing in basements, shear-wall installation, water-heater anchoring, and securing of chimneys (Heinz Center, 2000).

Zoning Ordinances That Better Communicate Risk

One of the more vexing problems facing policy makers after major catastrophes is whether to permit reconstruction in areas that have been damaged. As the response after Katrina demonstrated, there is usually strong political support for wanting to rebuild one's home in the same place where it was damaged or destroyed. Indeed, not to do so somehow seems to show a lack of empathy for those who have lived part, if not all their life, in the area and have family and social connections there; for many of them, nowhere else could be home (Vigdor, 2008).

An unfortunate tendency after disasters is not only permitting homes to be rebuilt in hazard-prone areas, but rebuilding the structures so as to remove all signs that might communicate to new and prospective residents the inherent risks posed by the location. Visitors to the Mississippi Gulf coast today, for example, will find few cues that would be indicative of the complete devastation that the area suffered from Hurricane Katrina in 2005: attractive mansions are once again strung along Route 90, where sandy beaches give little clue that this is perhaps the most hazard-prone section of coastline in the United States.

While policies that prohibit residents from rebuilding destroyed residences may be politically unviable, policies that guide reconstruction in such a way that allows new residents to make informed decisions about the real risks they face would seem far less controversial. This notion is implicitly recognized in FEMA's flood maps, which the agency is in the process

of updating. We urge that this education process be recognized more widely, not just for floods, but for hurricanes and earthquakes as well (GAO, 2008).

Conclusions

Recent disasters in the United States have provided empirical evidence supporting the *natural disaster syndrome*. Following Hurricane Katrina, many victims in Louisiana suffered severe losses from flooding because they had not mitigated their homes and did not have flood insurance to cover the resulting damage. As a result, there was an unprecedented level of federal disaster assistance to aid these victims.

There are many reasons why those in harm's way have not protected themselves against natural disasters. In this chapter we have highlighted behavioral considerations that include budgeting heuristics, short-term horizons, underestimation of risk, optimism, affective forecasting errors, learning failures, social norms and interdependencies, the Samaritan dilemma, and the politician's dilemma. All these effects limit people's interest and ability to invest in hazard mitigation measures.

The 2004, 2005, and 2008 hurricanes should have been a wake-up call in this regard, but instead they seemed to fall on deaf ears. The next series of major hurricanes, floods, or earthquakes are likely to be devastating ones.

In the case of New Orleans, inhabitants may have simply felt that they were fully protected by flood-control measures, such as the levees. Unfortunately, it is very likely that we will discover after future catastrophes that there are actually many similar situations of false perceived security in other highly exposed areas (Pinter, 2005). The question of the resiliency of our infrastructure to large-scale disasters is one that has not yet received the attention it deserves (Auerswald et al., 2006).

If we as a society are to commit ourselves to reducing future losses from natural disasters and to limiting government assistance after the event, then we need to engage the private and public sectors in creative partnerships. This requires well-enforced building codes and land-use regulations coupled with adequate insurance protection. Economic incentives that make these actions financially palatable to property owners should be provided in the form of long-term insurance policies and mitigation loans.

One may also want to think more about the type of disaster insurance that should be provided in those hazard-prone areas. It may be useful to consider the possibility of providing protection against all hazards under a long-term homeowner's insurance policy tied to the mortgage rather than continuing with the high volatility inherent in annual insurance contracts. These and related issues form the basis for future behavioral studies that may help us to develop more effective policy recommendations for reducing losses from future natural disasters.

Notes

1. A discussion of alternative flood reduction measures can be found in Laska (1991) and Federal Emergency Management Agency (1998).

2. Still, the U.S. Census Bureau estimates indicate that almost two years after the storm, by July 1, 2007, nearly half of these evacuees had yet to return to New Orleans (Vigdor, 2008).

3. Following a flood in August 1998 that damaged property in northern Vermont, FEMA found that 84% of the 1,549 homeowners suffering damage resided in Special Flood Hazard Areas but did not have insurance, even though 45% of those individuals were required to purchase this coverage (Tobin and Calfee, 2005).

4. For more details on the proposed long-term flood insurance policy, see Kunreuther and Michel-Kerjan (2010).

References

Auerswald, P., Branscomb, L., La Porte, T., and Michel-Kerjan, E. (2006). *Seeds of disasters, roots of response: How private action can reduce public vulnerability.* New York: Cambridge University Press.

Berg, R. (2009). *Hurricane Ike tropical cyclone report.* Miami: National Hurricane Center. Retrieved from http://www.nhc.noaa.gov/pdf/TCR-AL092008_Ike_3May10.pdf

Bowman, E., and Kunreuther, H. (1988). Post-Bhopal behavior in a chemical company. *Journal of Management Studies*, *25*, 387–402.

Brinkley, D. (2006). *The great deluge: Hurricane Katrina, New Orleans, and the Mississippi Gulf Coast.* New York: Harper Collins.

Buchanan, J. (1975). The Samaritan's dilemma. In E. Phelps (Ed.), *Altruism, morality and economic theory* (pp. 71–85). New York: Russell Sage Foundation.

Burby, R. J. (1991). *Sharing environmental risks: How to control governments' losses in natural disasters.* Boulder, CO: Westview.

———. (2006). Hurricane Katrina and the paradoxes of government disaster policy: Bringing about wise governmental decisions for hazardous areas. *Annals of the American Academy of Political and Social Science*, *604*, 171–191.

Burby, R. J., Bollens, S. J., Kaiser, E. J., Mullan, D., and Sheaffer, J. R. (1988). *Cities under water: A comparative evaluation of ten cities' efforts to manage*

floodplain land use. Boulder, CO: Institute of Behavioral Science, University of Colorado.

Camerer, C., and Kunreuther, H. (1989). Decision processes for low probability events: Policy implications. *Journal of Policy Analysis and Management*, *8*, 565–592.

Campbell, J. Y. (2006). Household finance. Working paper, Department of Economics, Harvard University.

Congressional Research Service. (2005). *Federal Stafford Act Disaster Assistance: Presidential declarations, eligible activities, and funding*. Washington, DC: Congressional Research Service, Library of Congress. Retrieved from http://www.au.af.mil/au/awc/awcgate/crs/rl33053.pdf

Dacy, D., and Kunreuther, H. (1968). *The economics of natural disasters*. New York: Free Press.

Federal Emergency Management Agency (FEMA). (1998). *Homeowner's guide to retrofitting: Six ways to prevent your home from flooding*. Washington, DC: Federal Emergency Management Agency.

Fritz, H. M., Blount, C. D., Swe Thwin, Moe Kyaw Thu, and Nyein Chan (2009). Cyclone Nargis Storm surge in Myanmar. *Nature Geoscience*, *2*, 448–449.

Gladwell, M. (2000). *The tipping point*, New York: Little Brown.

Goodnough, A. (2006, May 31). As hurricane season looms, states aim to scare. *New York Times*. Retrieved from http://www.nytimes.com/2006/05/31/us/31prepare.html?pagewanted=all

Government Accountability Office (GAO). (2008). *Flood insurance: FEMA's rate-setting process warrants attention*. GAO-09-12. Washington, DC: U.S. Government Accountability Office.

Heal, G., and Kunreuther, H. (2005). You can only die once: Interdependent security in an uncertainty world. In H. W. Richardson, P. Gordon, and J. E. Moore, II (Eds.), *The economic impacts of terrorist attacks* (pp. 35–56). Cheltenham, UK: Edward Elgar.

H. John Heinz III Center for Science, Economics and the Environment (Heinz Center). (2000). *The hidden costs of coastal hazards: Implications for risk assessment and mitigation*. Washington, DC: Island Press.

Hogarth, R., and Kunreuther, H. (1995). Decision making under ignorance: Arguing with yourself. *Journal of Risk and Uncertainty*, *10*, 15–36.

Huber, O., Wider, R., and Huber, O. W. (1997). Active information search and complete information presentation in naturalistic risky decision tasks. *Acta Psychologica*, *95*, 15–29.

Jaffee, D., Kunreuther, H., and Michel-Kerjan, E. (2008). *Long term insurance (LTI) for addressing catastrophe risk*. NBER Working Paper No. 14210. National Bureau of Economic Research.

Knabb, R. D., Rhome, J. R., and Brown, D. P. (2006). *Tropical cyclone report: Hurricane Katrina: 23–30 August 2005*. Miami: National Hurricane Center.

Kunreuther, H. (1996). Mitigating disaster losses through insurance. *Journal of Risk and Uncertainty*, *12*, 171–187.

———. (2006). Disaster mitigation and insurance: Learning from Katrina. *Annals of the American Academy of Political and Social Science*, *604*, 208–227.

Kunreuther, H., Ginsberg, R., Miller, L., Sagi, P., Slovic, P., Borkan. B., and Katz, N. (1978). *Disaster insurance protection: Public policy lessons*. New York: John Wiley and Sons.

Kunreuther, H., and Michel-Kerjan, E. (2006). *Some options for improving disaster mitigation: Developing new commercial opportunities for banks and Insurers*. Paper prepared for the Financial Services Roundtable. Washington, DC.

———. (2009). *At war with the weather*. Cambridge, MA: MIT Press.

———. (2010). Market and government failure in insuring and mitigating natural catastrophes: How long-term contracts can help. In W. Kern (Ed.), *The economics of natural and unnatural disasters* (pp. 9–38). Kalamazoo, MI: W. E. Upjohn Institute for Employment Research.

Kunreuther, H., Onculer, A., and Slovic, P. (1998). Time insensitivity for protective measures. *Journal of Risk and Uncertainty*, *16*, 279–299.

Kunreuther, H., and Pauly, M. (2005). Terrorism losses and all-perils insurance. *Journal of Insurance Regulation*, *23*, 3–19.

———. (2006). Rules rather than discretion: Lessons from Hurricane Katrina. *Journal of Risk and Uncertainty*, *33*(1–2), 101–116.

Kydland, F., and Prescott, E. (1977). Rules rather than discretion: The inconsistency of optimal plans. *Journal of Political Economy*, *85*, 473–91.

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics*, *112*, 443–477.

Laska, S. B. (1991). *Floodproof retrofitting: Homeowner self-protective behavior*. Boulder, CO: Institute of Behavioral Science, University of Colorado.

Lerner, J., Gonzalez, R., Small, D., and Fischhoff, B. (2003). Effects of fear and anger on perceived risks of terrorism: A national field experiment. *Psychological Science*, *14*(2), 144–150.

Loewenstein, G., O'Donoghue, T., and Rabin. M. (2003). Projection bias in predicting future utility. *Quarterly Journal of Economics*, *118*, 1209–1248.

Loewenstein, G., and Prelec, D. (1992). Anomalies in intertemporal choice: Evidence and an interpretation. *Quarterly Journal of Economics*, *107*(2), 573–597.

Magat, W., Viscusi, K. W., and Huber, J. (1987). Risk-dollar tradeoffs, risk perceptions, and consumer behavior. In W. Viscusi and W. Magat (Eds.), *Learning about risk* (pp. 83–97). Cambridge, MA: Harvard University Press.

McClelland, G., Schulze, W., and Coursey, D. (1993). Insurance for low-probability hazards: A bimodal

response to unlikely events. *Journal of Risk and Uncertainty*, *7*, 95–116.

Meyer, R. (2006). Why we under prepare for hazards. In R. J. Daniels, D. F. Kettl, and H. Kunreuther (Eds.), *On risk and disaster: Lessons from Hurricane Katrina* (pp. 153–174). Philadelphia: University of Pennsylvania Press.

Meyer, R. J., Zhao, S., and Han, J. K. (2007). *Biases in valuation and usage of innovative product features.* Working paper, Department of Marketing, the Wharton School, University of Pennsylvania.

Michel-Kerjan, E. (2008a). Disasters and public policy: Can market lessons help address government failures. Proceedings of the 99th National Tax Association Conference, Boston, MA. *National Tax Association Proceedings*, 179–187. Retrieved from http://www.ntanet.org/images/stories/pdf/proceedings/06/023.pdf

———. (2008b). Toward a new risk architecture: The question of catastrophe risk calculus. *Social Research*, *75*(3), 819–854.

Michel-Kerjan, E, and Kousky, C. (2010). Come rain or shine: Evidence on flood insurance purchases in Florida. *Journal of Risk and Insurance*, *77*(2), 369–397.

Moss, D. (2002). *When all else fails: Government as the ultimate risk manager.* Cambridge, MA: Harvard University Press.

Oberholzer-Gee, F. (1998). *Learning to bear the unbearable: Towards an explanation of risk ignorance.* Unpublished manuscript, Wharton School, University of Pennsylvania.

Palm, R., Hodgson, M., Blanchard, R. D., and Lyons, D. (1990). *Earthquake insurance in California: Environmental policy and individual decision making.* Boulder, CO: Westview Press.

Pielke, R., Gratz, J., Landsea, C., Collins, D., Sanders, M., and Musulin, R. (2008). Normalized hurricane damage in the United States: 1870–2005. *Natural Hazards Review*, *9*(1), 29–42.

Pinter, N. (2005). One step forward, two steps back on U.S. floodplains. *Science*, *308*(5719), 207–208.

Reeves, A. (2004, May 12). Plucking votes from disasters. *Los Angeles Times.* Retrieved from http://articles.latimes.com/2004/may/12/opinion/oe-reeves12

———. (2005). *Political disaster? Electoral politics and presidential disaster declarations.* Manuscript in progress. Kennedy School of Government, Harvard University, Cambridge, MA.

Schelling, T. (1978). *Micromotives and Macrobehavior.* New York: Norton.

Steinberg, T. (2000). *Acts of God: The unnatural history of natural disaster in America*. New York: Oxford University Press.

Stern Review. (2006). The economics of climate change. H. M. Treasury. Retrieved from http://webarchive.nationalarchives.gov.uk/+/http:/www.hm-treasury.gov.uk/sternreview_index.htm

Sunstein, C. R. (2006). Deliberating groups vs. prediction markets (or Hayek's challenge to Habermas). *Episteme: A Journal of Social Epistemology*, *3*, 192–213.

Thaler, R. (1999). Mental accounting matters. *Journal of Behavioral Decision Making*, *12*, 183–206.

Tobin, R., and Calfee, C. (2005). *The National Flood Insurance Program's mandatory purchase requirement: Policies, processes, and stakeholders.* Washington, D.C.: American Institutes for Research.

Trope, Y., and Liberman, N. (2003). Temporal construal. *Psychological Review*, *110*(3), 403–421.

Tversky, A., and Shafir, E. (1992). Choice under conflict: The dynamics of deferred decision. *Psychological Science*, *3*(6), 358–361.

Vigdor, J. (2008). The economic aftermath of Hurricane Katrina. *Journal of Economic Perspectives*, *22*(4), 135–54.

White, G. F. (1975). *Flood hazard in the United States: A research assessment.* Boulder: Institute of Behavioral Science, University of Colorado.

Wilson, T. D., and Gilbert, D. T. (2003). Affective forecasting. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 35, pp. 345–411). San Diego, CA: Academic Press.

CHAPTER 24

Decisions by Default

ERIC J. JOHNSON

DANIEL G. GOLDSTEIN

Imagine that it is the December after a presidential election in the United States. The president-elect is interviewing to fill cabinet positions. One potential appointee makes an offer that seems too good to refuse: "Appoint me and I will let you in on a secret policy-setting strategy that has been shown to increase savings rates, save lives by improving voluntary efforts, and change how people manage risk." Even better, the potential appointee adds, "It costs very little and works even if people do nothing. In fact, it *depends* upon people doing nothing." What is the secret strategy? The proper setting of *no-action defaults*.

Default options are those assigned to people who do not make an active decision.[1] For example, in the United States, people are by default *not* organ donors and, until recently, did *not* contribute to their retirement plans. However, when purchasing new automobiles, they *do* receive air bags by default; customers need not ask for them. How strong are the effects of defaults on choices? Do defaults influence us and why? Drawing on a variety of policy domains, we will illustrate the power of default effects and explore the possible psychological mechanisms that underlie them. With an eye on questions of ethics, obligations, and effectiveness, we will evaluate policies that have been proposed for setting defaults. Our goal is to move defaults from the chest of secret strategies to the library of mechanisms by which policies can impact behavior.

Default Effects on Choices: Case Studies

INSURANCE DECISIONS

By assigning similar groups of people to different policies, governments, companies, and public agencies sometimes run inadvertent "natural experiments" that allow the effects of defaults to be estimated. Default effects have been found in the choice of health-care plans (Samuelson and Zeckhauser, 1988), and the choice between privacy policies online (Bellman, Johnson, and Lohse, 2001). People have a strong tendency to choose the default option to which they were assigned, even when this assignment is random. When the stakes are low, we might expect people to stick with such a randomly assigned alternative. However, defaults affect choices even when the stakes are high. In the 1990s, the states of New Jersey and Pennsylvania gave consumers of auto insurance a choice between a more expensive plan, which provided the right to sue for "pain and suffering," and a significantly less expensive plan, which covered the medical costs of the insured but removed the right to sue. New Jersey drivers were given the limited right to sue by default, whereas Pennsylvania drivers had the opposite default, the full right to sue. Surprisingly, 21% of New Jersey drivers "preferred" the more expensive full right to sue, whereas in Pennsylvania over 70% of drivers "preferred" the less expensive plan. A psychological study in which people were assigned one of the two plans by default confirmed this: the full right to sue was chosen 53% of the time when it was the default, but only 23% of the time when it was not (Johnson et al., 1993). One estimate (Goldstein et al., 2008) is that the choice of defaults in Pennsylvania resulted in $140 million a year in additional insurance purchases, or a total of $2 billion since 1991.

ORGAN DONATION

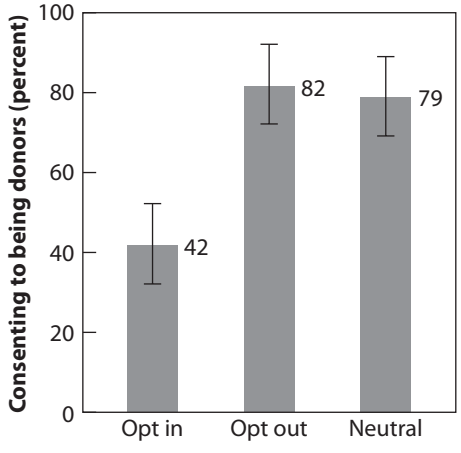
Over the course of the last two decades, a number of European countries have been running similar natural experiments with organ donation. Different countries have chosen different defaults for membership in organ donor pools. *Opt-in countries* require explicit consent to become a donor, while *opt-out countries* presume consent, requiring an active step to leave the pool.

We (Johnson and Goldstein, 2003) examined the role of defaults using an online experiment. We asked 161 respondents whether they would be donors using

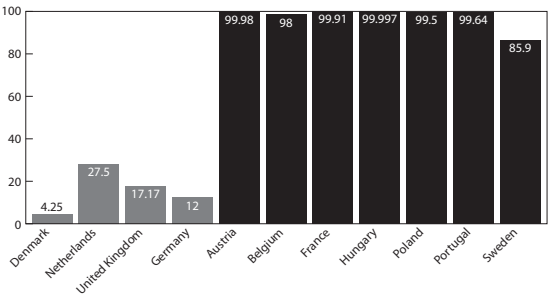
one of three questions. In the opt-in condition, participants were told to assume that they had just moved to a new state where the default was to *not be* an organ donor; they were given a choice to confirm or change that status. The opt-out condition was identical, except the default was *to be* a donor. The third, neutral, condition simply required them to choose with no prior default. The effort needed to complete the experiment was the same in all conditions: respondents could change their choice with only a mouse click.

The randomly assigned default option had a dramatic impact, with revealed donation rates being about twice as high when opting out as when opting in. As can be seen in figure 24.1, the opt-out condition did not differ significantly from the neutral condition, which required a choice without a default option. Only the opt-in condition, the current practice in the United States, was significantly lower.

Because there are many factors that might produce such effects in the real world, we examined the rate of agreement to become a donor across European countries with explicit- or presumed-consent laws. From data reported in Gäbel (2002), which we supplemented by contacting central registries for several countries, we estimated the effective consent rate, that is, the number of people who had opted in (in explicit-consent countries) or the number who had not opted out (in presumed-consent countries). If preferences concerning organ donation are strong, defaults should have little or no effect. However, as can be seen in figure 24.2, defaults make a large difference, with the four opt-in countries on the left having lower rates than the seven opt-out countries on the right. The result is surprisingly strong: the two distributions have no overlap and nearly 60 percentage points separate the highest opt-in and the lowest opt-out countries. We suspect these effects are larger than those in our questionnaire because the cost of changing from the default is higher (e.g., filling out forms, making phone calls, or sending mail to change status).



24.1. Effective consent rates, online experiment, by default.

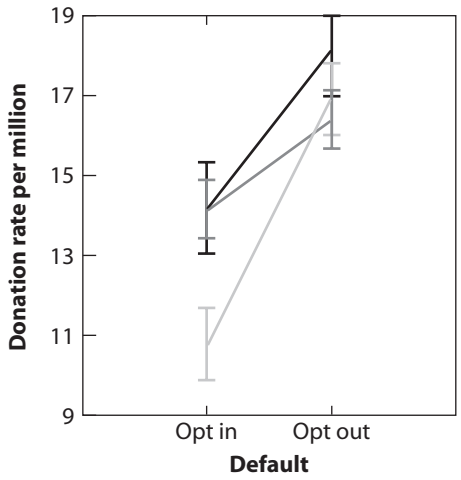


24.2. Effective consent rates by country; the four gray bars indicate explicit consent (opt in) and the seven black bars indicate presumed consent (opt out).

Some opt-in countries have tried hard to increase donations: the Netherlands, upon creating its national registry, launched an extensive educational campaign and a mass mailing of more than 12 million letters (in a country of 15.8 million) asking citizens to register, but this failed to substantially change the effective consent rate (Oz et al., 2003).

Could changes in defaults have an effect on the actual number of donations in a country? Using a time series of data from 1991 to 2001, we examined the actual number of caderveric donations made per million on a slightly larger list of countries. We used a regression analysis that controlled for differences in countries' propensity toward donation, transplant infrastructure, educational level, and religion; all variables known to affect donation rates (Gimbel et al., 2003).

While there are no differences across years, there is a strong effect of the default: figure 24.3 shows that when donation is the default, there is a significant ($p < .02$) increase in donation, increasing from 14.1 to 16.4 per million, a 16.3% increase. Using similar techniques but looking only at 1999 for a broader set of countries, including many more from Eastern Europe, Gimbel et al. (2003) reported an increase from 10.8 to 16.9, a 56.5% increase (fig. 24.3). An alternate specification of the time series analysis by Abadie and Gay (2006) showed a larger increase that they claim, if applied to the United States, would eliminate the shortage in certain categories of organs.



24.3. Estimated mean donation rate, 1991–2001, in donors per million as a function of default (opt in vs. opt out) for two alternate time series analyses (Johnson and Goldstein, 2003, dark gray line; Abadie and Gay, 2006, black line), and one cross-sectional analysis (Gimbel et al., 2003, light gray line).

RETIREMENT SAVINGS

A third example involves contributions to retirement savings plans, perhaps the most important financial decision facing most Americans. Many workers are covered by a defined contribution plan, a 401(k) plan for example, in which people elect to save between 0% and 12% of their income for retirement. The plans are attractive: the contributions are in pretax dollars, the money compounds free of tax, and an employer often matches the first 6%. Consistent with observations that Americans are not saving sufficiently toward retirement, many people initially contribute the default, that is, nothing. However, when the default was altered, the result was a marked increase in savings (Madrian and Shea, 2001). One firm raised the default contribution from 0% to 3% and saw the percentage of new employees saving *anything* toward retirement rise from 31% to 86%. However, the effect was almost too powerful: increasing the default to 3% surprisingly decreased the number of people electing to save more than 3%. This result has been replicated in several firms (Choi et al., 2001), raising questions about what default is optimal. A review of the effectiveness of automatic enrollment options in Save More Tomorrow plans (Benartzi, Peleg, and Thaler, 2009) finds that enrollment rates are about 25% when people have to opt in to automatic increases of their savings, but 84% when they must opt out. With economic stakes as large as these, it seems unlikely that default effects are attributable to "rational inaction" (Samuelson and Zeckhauser, 1988).

INTERNET PRIVACY POLICIES

When consumers sign up for a new Internet service, install a new piece of software, or order merchandise electronically, they are often presented with choices accompanied by a long and complex privacy policy statement. These policies outline the rights that the consumer and the firm agree to in conducting the transaction. These transactions often have default options. For example, the search site Google asks for explicit permission (opt in) to share a user's data, but once that is given, Google assumes permission to extend that use (opt out). Such distinctions were a major point of contention between the United States and the European Union privacy policies, leading to what is termed the safe-harbor agreement harmonizing the two standards (Bellman et al., 2004).

We conducted a web-based experiment examining the willingness of visitors to a website to be contacted with solicitations for surveys (Bellman et al., 2001). Asking people to opt in to these paid surveys resulted in a 48.2% participation. When people had to opt out to not participate, 96.3% agreed to participate.

SEX EDUCATION IN KANSAS

Sex education in schools has been a source of controversy, and many localities have allowed parents to opt their children out of this instruction. However, several states, most recently Kansas, have changed the default. After significant controversy, including a tie vote, the Kansas State Board of Education joined three other states (Arizona, Nevada, and Utah) in requiring explicit permission for attendance in sex-education classes.[2]

MILITARY RECRUITMENT AND THE NO CHILD LEFT BEHIND ACT

An obscure element of the 2001 No Child Left Behind Act is the requirement that high schools give students' names, addresses, and telephone numbers to military recruiters unless a parent explicitly informs the school district. The opt-out nature of consent here is explicitly stated in the law, perhaps in response to some districts banning recruiters from campus in response to the Don't Ask-Don't Tell policy.[3]

The opt-out requirement has been both controversial and effective. In most school districts the majority of permission slips are not returned, which is the equivalent of granting permission. Few parents

explicitly object to the release of information. Fairport, a district near Rochester, New York, decided to change the default. In this district, only 5% of the parents gave explicit permission. Fairport attracted national attention and the threat of losing all federal funding for the district if the opt-in policy stayed in place.

DO NOT CALL REGISTRY

The National Do Not Call Registry is an apparent example of defaults not mattering. It has enrolled over 149 million households, almost all using a website. The bill had strong support in the Senate with a unanimous vote and the House by a 418–7 margin. It is prototypical of the idea of Notice and Consent, which has guided Federal Trade Commission Policy for the last twenty-five years (Center for Democracy and Technology, 2009), and provides partial protection against telemarketing calls. Exceptions are made for surveys, political speech (including robocalls), not-for-profit organizations, and firms that have an existing relationship with the consumer. Violating the Do Not Call Registry can result in significant fines for a company. Such a significant enrollment is impressive, yet one has to wonder, How many of the remaining people would really have opted in to receive telemarketing calls?

The Causes of Default Effects

Why do defaults make such a difference? While effects this strong and robust are likely to have several causes, three broad categories suggest themselves: effort, implied endorsement, and loss aversion. The causes of default effects should be of interest not just to psychologists. Policy makers, corporations, and marketers need to manage their defaults, and potential interventions are both suggested and justified by an understanding of the causal mechanisms behind defaults. Similarly, the choice of interventions to minimize or harness default effects depends upon their causes. If in one context, default effects are due to effort, effort reduction is the suggested treatment and selection of the default should match a best guess of what the choice would be if there were no default. Other causes, for example, that the default is interpreted to be a recommendation, have different implications for management and intervention.

Effort

In economics, transaction costs are a cause of market failures. They represent a source of friction that prevents convergence to equilibrium because some trades that would be required to reach equilibrium become too expensive to perform. Similarly, default effects may be seen as a kind of market failure due to the effort required to register a choice. If the cost of registering one's true preferences is high, then default options may be selected even when they would be rejected under frictionless choice. Such effects could well be classically rational if the benefits of expressing the preference are outweighed by the costs. Recall the insurance example of the drivers' right to sue. To make a decision, one would have to take the time to read a complex statement, decide what is best given personal circumstances, fill out the form, and then hunt down a stamp and remember to drop the envelope into the mailbox. This may not be rational inaction: while this list of activities appears onerous, it seems unlikely for most people that the required actions offset the annual $300 savings available to those who might pick the limited tort policy.

A second type of effort is that involved in forming a preference. One tenet in psychological decision-making research is that some preferences are not formed until a decision situation is encountered (Fischhoff, 1991; Payne, Bettman, and Johnson, 1992; Slovic, 1995). If deciding how we feel is effortful, this sort of just-in-time preference construction saves effort by avoiding tough choices about situations that may never be encountered. When eventually faced with difficult decisions, those who have not formed preferences can continue to avoid costs by adopting the default option as their preference. There is a large literature on preference construction that serves to suggest conditions under which default effects may be of particular concern. For example, people who are less experienced with a given attribute are more likely to exhibit some kinds of context effects due to preference construction (Payne, Bettman, and Schkade, 1999).

STRATEGIC INCREASES IN EFFORT

Those who set defaults often have the power to make switching away from them difficult, which will only amplify default effects. While the first author was being interviewed on National Public Radio's *Marketplace*, he had the opportunity to listen in as the host attempted to opt out of Verizon's choice to share his calling records with third parties. While a company spokesperson attributed the long opt-out process to ensuring that the correct caller was identified, it is hard to see why the company would need his phone number, the first 13 digits of the account number (which were identical to the phone number and read back to him slowly by an unpleasant automated voice), his spoken name, his spoken address, his

spoken name (a second time), and his agreement to "place a restriction upon his account" to ensure that he was who he said he was. More likely, this could be characterized as raising switching costs strategically.

DECISION EFFORT IN ECONOMICS AND PSYCHOLOGY

When does effort justify accepting a default? We need to make a distinction between rational inaction and regrettable inaction. Effort, in standard economics, is treated just like a transaction cost: if the benefit from expending the effort is less than or equal to the benefits of thinking, then the effort is made. Thinking is not free in standard economics, but its cost never exceeds the benefit that it generates. Thus, cognition is not a consideration because it simply has a net cost of zero (Gabaix et al., 2006). In contrast, it is clear there are cases in which far too little effort is expended given the stakes. One particularly compelling example involves decisions across time, such as cases in which expending effort in the present would prevent greater future effort or produce substantial future benefits. In these cases, excessive discounting of future benefits may contribute to increased default taking (Laibson, 1997; O'Donoghue and Rabin, 2001; Zauberman, 2003). This suggests an intriguing application of what we know about intertemporal choice and the conditions in which perceived effort might lead to increased default taking.

DEFAULTS AS AN EFFORT TAX

Defaults can impose costs upon decision makers. Consider the National Do Not Call Registry, which makes opting in difficult in at least two ways. First, access to the Internet is required to register. Perhaps most dramatically, the law contained an explicit shift in defaults. All phone numbers, by law, drop off the list five years after their registration.

The Federal Trade Commission (FTC) justified that this reenrollment is simple: "It is incredibly quick and easy to do," Lydia Parnes, director of the FTC's bureau of consumer protection, said in an interview with the Associated Press (Kerr, 2007). "It was so easy for people to sign up in the first instance. It will be just as easy for them to re-up." Yet, we doubt that this is welfare maximizing for the participants. In essence the FTC is imposing an asymmetric cost, penalizing the majority who would want to maintain their status and replacing it with an effort savings for those who would otherwise need to opt in. What may seem to be minor costs at the individual level are much larger in the aggregate. Assume that a significant majority of people (say 90%) would want to continue to opt out. Obviously calculating such costs may be a foolhardy exercise, but even if the "incredibly quick" task takes five minutes, then this imposes over 1,000 person years of time reaffirming a stable preference even considering the savings for those who would prefer to hear from telemarketers.[4]

As this example suggests, defaults save effort only for those whose preferences coincide with the default option. Heeding effort considerations, one plausible principle for decision makers would be to choose defaults that mimic majority preferences (Sunstein and Thaler, 2003).

Implied Endorsement

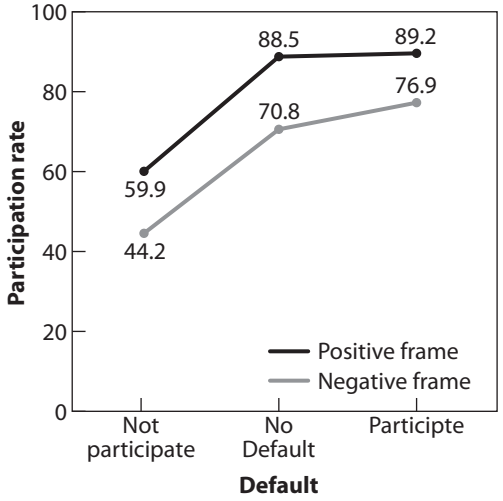
In the case of policy defaults, such as for organ donor status or pension plan membership, McKenzie, Liersch, and Finkelstein (2006) suggest that people interpret defaults as a recommended course of action set out by policy makers. Consider the Kansas Sex Education case. A parent not knowledgeable of the content of a sex ed course may be unsure of their preferences for their child. The course may or may not fit their values, and it is not simple to predict how the course might be inappropriate: a serious treatment of abstinence or alternative lifestyles in the classroom may be deemed inappropriate by different parents. An important cue may be whether the board of education thinks the curriculum is suitable for most people (as might be signified by an opt-out policy) or maybe just a minority (as signified by an opt-in policy). McKenzie, Liersch, and Finkelstein argue that defaults can contain information about what the policy maker thinks is the right thing to do.

Sunstein and Thaler (2003) propose that the default selected by policy makers might be interpreted as an indication of what the majority chooses and that following a simple heuristic of imitation could lead to its widespread adoption (Henrich et al., 2001). In a marketplace context, Brown and Krishna (2004) posit that defaults set by marketers may be perceived as suggestions, and in the case of suspicious vendors, as manipulation attempts. Their experiments find that default effects are diminished or even backfire when consumers become sufficiently skeptical. The view of defaults as endorsements does not portray the selection of defaults as arising from cognitive limitations; on the contrary, it suggests that agents react to defaults with a kind of developed intelligence or "marketplace metacognition" (Wright, 2002). To the extent this occurs, defaults might backfire, producing reactance.

Loss Aversion and Reference Dependence

Every nontrivial choice involves a trade-off, surrendering one thing to gain another. For example, consider

retirement savings. To start contributing to a retirement plan, an employee must reduce present take-home pay to gain future income; the trade-off seems to be between a present loss and a future gain. However, if the employee were auto-enrolled in the retirement plan, then the choice to opt out might seem like that between a future loss (reduced retirement income) and a present gain (a bigger paycheck). Formally, of course, these decisions are identical, but psychologically they are not. Loss aversion suggests that what is given up will have more impact upon choices than what is gained (Köszegi and Rabin, 2006; Tversky and Kahneman, 1991). The change in reference points caused by defaults may result in a fairly remarkable change in the attractiveness of the two options. In the first case, the forgone income, the loss, has a greater impact upon choice; in the second, the lost future income is the loss and has a greater impact upon choice. This combination of shifting reference and loss aversion has been thought of as one of the major causes of default effects (Johnson et al., 1993; Samuelson and Zeckhauser, 1988). In addition, default effects bear a similarity to endowment effects, in which the mere possession of an object increases its value relative to an amount of money. Defaults can be seen as instant endowments that cause an increase in value. Together, reference dependence and loss aversion seem to provide a shift in perspective that can cause an option to seem more attractive.



24.4. Default effects for framing and prechecking of form boxes.

Empirical Evidence

Given the size and robustness of default effects, we think it is unlikely that any of these three causes will emerge as the primary one. All three play roles in different situations, and it seems quite possible that the strength of each differs according to context.

While there is a need for further research examining the contribution of each cause, some guidelines may prove useful. Effort is most likely to have an effect when recording a preference is difficult or when preferences are unknown. As is the case when dealing with lengthy insurance forms or robotic telephone instructions, the costs of switching can be real. However, with increasing consumer demand for user-friendly, web-based service platforms, the role of effort may be diminishing. As mentioned earlier, the default setting by the Kansas Board of Education might be a case in which implied endorsement looms large. Some websites, such as Dell for example, use explicit endorsement actually separating the default and recommended options. Finally, loss aversion probably plays a role, although this role has largely been assumed and not empirically demonstrated.

The relative role of two sources of default effects in a realistic scenario was examined in a web-based study that separately manipulated whether or not the default was already checked when presented to the decision maker and whether or not the language was framed positively (get further surveys) or as a loss (do *not* get further surveys) (Bellman, Johnson, and Lohse, 2001). The prechecking of the box would seem to combine a minimal-effort (since it is a web page) situation and an implied-endorsement situation. Supporting our argument that default effects have multiple causes, the rather weak linguistic manipulation had about a 16% difference in sign-up rates. The prechecked box produced a change of about 30%, and interestingly, there was no interaction, suggesting that these were additive effects (fig. 24.4).

Dinner et al. (2011) examined the effect of defaults upon decisions, explicitly looking for evidence supporting a loss-aversion account. They asked participants what kind of lightbulb they would prefer to have installed in a new renovation and manipulated the default as either a standard incandescent bulb or a more energy efficient, but more expensive, compact florescent bulb. As expected, they found a significant advantage with the default: 42% of the respondents chose the incandescent bulb when it was the default, but only 20% did so when it was not. They also asked participants to provide a thought listing as they made their choice. Consistent with the idea that a default changes the way that options are viewed, they found that the default option was mentioned earlier and more positively than the alternative. They argue that this is consistent with loss-aversion explanations of defaults and do not find support, in this particular case, for either effort or implied endorsement.

Managing Defaults

Defaults and their effects are ubiquitous; however, knowledge about them is not. As a result, managers and policy makers often neglect the topic of defaults altogether. Recall how a month after the introduction of the much-hyped iPhone, customers started receiving very long itemized phone bills; in some cases the bills were hundreds of pages long. Almost all iPhone customers received these environmentally unfriendly bills, and many were surprised. How could this come about? AT&T, apparently in the rush to go to market, had included a prechecked question asking customers whether they wanted detailed billing information. In this case, default neglect led to wasted paper and postage. Why did Pennsylvania and New Jersey end up with opposing default policies for insurance? It has been argued that Pennsylvania's trial lawyers made the change at the last minute, and that the governor did not think that the default would matter. We now know that he was wrong.

In the movie *The Year of Living Dangerously*, the translator and guide Billy Kwan borrows a line from Tolstoy and asks, "What then must we do?" While both Tolstoy and Kwan were talking about poverty, the question applies to defaults in public policy as well. After all, when designing choices, there is not a no-default option. What tools do policy makers have at their disposal? We present some here, with no claims of a complete list (for a more detailed inventory and analysis, see Goldstein et al., 2008).

Forced Choice or Default?

A policy maker might attempt to prevent people from *not* making a choice—a situation that we call *forced choice* (or mandated choice). This policy has appeal in cases where a policy maker is unclear about what option is best for most people (Goldstein et al., 2008).

In our research on organ donation, for example, we constructed a web page that required respondents to make a response before they could move on. Recall that this resulted in reported preferences that mimicked those of the opt-out group (see fig. 24.1). This is an important result because it suggests that the opt-in default is biased relative to no default, and that the opt-out default is closer to most people's expressed wishes.[5]

Forced choice does come with a cost: the cost of making a decision. Although people find the thought of donating their own organs an aversive thought, defaults allow people to not make choices. Field experience with forced choice is instructive. The state of Virginia adopted a policy of asking people to make a choice about organ donation, but over 24% refused to report a preference (Klassen and Klassen, 1996). Results like this one suggest that forming a preference may be costly in both cognitive and emotional terms. If this is the case, when the majority of people have a preference to donate, a default would spare this majority the effort of making an onerous choice. Defaults not only make a difference in what is chosen, they can also make decisions easier. Mandated choice, in contrast, forces these costs on all.

The alternative to forced choice is the use of a default. However, defaults come in many varieties. In the next section we will summarize these and provide some guidelines for their use.

Mass Defaults

In many cases, it is desirable to have defaults that are the same for all people. Various laws might dictate that all people be presented with the same set of options expressed in the same way. In these cases, the only alternative to forced choice is a "mass default" policy.

BENIGN DEFAULTS

If a no-default option is not viable, a reasonable alternative would be a benign default, that is, the option that would be selected had no default been present. The appeal of this principle is simple: it appears to "do no harm" because one might assume that choices might be unaffected by the default. However, this option is less attractive than it might initially appear. One reason is that the default option might be selected too often, in part because it is the default. Recall that setting defaults for retirement savings increased savings overall but that the default of 3% reduced the number of people saving at a higher rate. A second, related issue is consumer heterogeneity: people differ in what constitutes for them a good decision. To illustrate, consider the example of automobile air bags. They do save lives, and the law requires their installation in all automobiles sold in the United States. However, they do not universally increase utility for all people. They can cause injury, or even death to small-stature people, particularly women and children. The benefits accrue more to those who are most likely to get in accidents, such as inebriated or careless drivers. Thus, while air bags do save lives, they also do not universally improve welfare for everyone. As a universal policy, this might cause some drivers harm and ignores the possibility of taking advantage of what is known about specific decision-makers. When purchasing automobiles, a 90-pound, 5-foot, 1-inch woman might be presented with a different default than, say, a 280-pound, 6-foot, 4-inch man. There are many applications of defaults where there are significant differences in needs.

Another simple example would be retirement investing, where an important decision is the allocation of funds between investments. Here, a one-size-fits-all default may yield better outcomes than a dominated option (such as leaving the money uninvested). This would suggest that a benign default would be good. But this would be inferior to what we will call a personalized default. Thus benign defaults seem most appropriate (1) when there is little heterogeneity in preferences, and (2) when most people might not make good decisions in the absence of defaults. The main benefit of benign defaults would be effort savings on the part of decision makers, and a potential increase in the quality of decisions.

RANDOM DEFAULTS

An alternative to identifying a good 'one-size fits all' benign default would be to randomly assign people to options. At first blush this may seem rather cavalier: surely people might do better left to their own devices. Yet policy makers have used such defaults. If a senior did not make an active choice for Medicare Part D prescription drug coverage, they were randomly assigned to a provider and plan. This reflects, perhaps, the uncertainty of officials in determining how to choose benign defaults. While they have been criticized, they do have one advantage: randomized defaults can help an organization learn what different people might choose and thus set better defaults in the future. By monitoring how people change their choices, policy makers can observe which defaults are the most popular and which ones lead people to bad choices. In the absence of any information suggesting better defaults, the use of random defaults can be a useful, albeit temporary, tool as part of an ongoing process of learning how to be of greater service to people.

Personalized Defaults

Under personalized defaults, knowledge about individuals is used to custom-tailor their defaults. Many firms, especially those that interact with customers online, are actively experimenting with personalized defaults (Goldstein et al., 2008). Since the interaction between policy makers and the public is moving online, we expect that personalized defaults will play a leading role in the future of policy design. Two basic personalized defaults are persistent defaults and smart defaults.

PERSISTENT DEFAULTS

A persistent defaults policy takes as a default that which the person chose last time. For instance, when a customer specifies to an airline that they prefer aisle seats, the airline may continue to assume the passenger wants an aisle seat unless they actively indicate otherwise. As we have seen with the example of the Do Not Call registry, some policy defaults revert rather than persist, to the possible displeasure of those who actively joined the registry. A persistent default is useful when one's past choices are likely predictors of current choices.

SMART DEFAULTS

A smart defaults policy uses individual measurements, such as a person's demographic or geographic profile, to assign a default that is likely to be suitable. In the case of allocating retirement savings, default options might be selected in consideration of the age of the participant: an older employee might be presented with a less risky allocation by default in comparison to a younger one. To return to our air-bag example, we would want to customize the deployment rate of the air bag to fit the stature of the driver. In fact, since 2005, such "smart" air bags have been issued in new automobiles. Such knowledge of what consumers would want, is, after all, one of the major activities of the marketing function; it may arguably be a function of public policy as well. Of course, individualized predictions are less than perfect, and gathering the information needed to customize defaults can have its costs. However, the cost of not customizing is also quite real, and if one believes that defaults have strong effects, one might want to ensure that they are set as intelligently as possible. In addition, correct classifications have the benefit of effort savings for decision makers.

Costs, Benefits, and Efficient Defaults

Every choice of default has costs and benefits. To help illustrate this, we have characterized those costs in table 24.1 using the data from our study on organ donation (Johnson and Goldstein, 2003). In particular, we drew on the no-default condition (to identify the intended categorization that is present without defaults) and the data from the opt-in and opt-out conditions to identify the effect of each default, assuming that the observed default effect affected intended donors and nondonors equally.

This table suggests three observations. First, almost every public policy has a no-action default, and the wise selection of defaults entails a balance between these costs. Using data from the study, we can see how the defaults affect choices. If the forced-choice condition reflects what people really want to do, it

Table 24.1 Relationship between defaults, categorization, and types of errors

	Intended categorization (no default)	
Realized categorization (opt in)	Donor (79%)	Not donor (21%)
Donor (42%)	Correct classification 33%	Incorrect classification, potential for indignation, negative publicity 8%
Not donor (58%)	Incorrect classification; potential lives saved forgone 45.8%	Correct classification 12%

appears that 45% of the population consists of potential donors who are misclassified in the opt-in frame currently used in the United States. These are people who would donate if forced to make decisions. These are quite rough estimates of the realized categories, since the procedures used to obtain agreement are often complex and depend upon the agreement of family members, and default effects are probably not independent of no-default choices. However, the point remains that default choices have real costs and that using the wrong default imposes very real opportunity costs.

Second, the idea that preferences are constructed provides an important alternative to the view that incentives are required to change behavior. For organ donation, there is a widespread belief in some circles that organ shortages are due to the lack of a market with incentives for the provision of live donations. Becker and Elias (2007) estimate that the shortage of living donations of kidneys could be overcome with a market price of $5,000. However, the available data suggest that most Americans approve of organ donation in the abstract but that many fewer have decided to become donors. Our diagnosis is different: we believe that many people have not made a decision, and the resulting prescription suggests that defaults will have a significant role in determining their status.

Finally, there is another cost, which is not considered in table 24.1—the cost of making a decision. In table 24.1, 45% (33% + 12%) of the population would be correctly classified if they do not change the default. However, an almost identical proportion, 45%, is misclassified because they do not make an active decision. The imposition of an opt-in requirement imposes an effort upon them that they do not undertake, costing many lives. A no-default option would impose the cost of making and registering a decision on the entire population. Finally, the opt-out default comes closer to what is observed in the forced-choice condition and imposes the cost of making an active choice on a smaller group, the 21% who do not want to be donors.

Defaults Are Cheap Compared to the Alternatives

Defaults are very efficient ways of changing behavior. To continue with the organ-donation example, recall that the Netherlands had attempted a large-scale mailing and public service campaign to increase the size of the donor pool. Although effective when compared to those of other countries, its effect was smaller than those resulting from changing the default. Since persuasive advertising *is* expensive, we would argue that there are many cases in which switching defaults is a much-lower-cost alternative. Thus, when compared to economic incentives or extensive educational or persuasive campaigns designed to influence people to make active decisions, changing a default can be an attractive alternative.

Notes

This research has been supported by grant SES-0352062 from the National Science Foundation, National Institute for Aging Grant 5R01AG027934, and the preparation of this manuscript has been aided by a grant from the Russell Sage Foundation to the first author. We thank participants at the Princeton Conference on the Behavioral Foundations of Policy and two reviewers for helpful comments.

1. Defaults are known by different names in different domains. The medical literature speaks of presumed versus explicit consent. In the domain of privacy policy and marketing, policies are classified as either *opt in* or *opt out*. In financial services, an action taken in the absence of a decision is sometimes referred to as a *negative election*.

2. Our thanks to Nicholas Epley for sharing this example.

3. Our thanks to Eldar Shafir for bringing this example to our attention.

4. Simply, this is the cost to those who do not want

phone calls (.9*149 million*5 minutes) minus the savings to those who do (.1*149 million*5 minutes) or 1,056.9 person years. The Do-Not-Call Improvement Act of 2007 retained prior selections but also kept the do-call default.

5. Subsequent research using similar questions and respondents shows two things that have changed since the 2002 execution of this study. First, consistent with national polling data, the number of people willing to donate has increased to more than 50%. Second, and perhaps as a result, the no-default condition has moved closer to the middle of the opt-in and opt-out defaults. Still, the fact remains that the majority of people in the no-default condition would donate.

References

Abadie, A., and Gay, S. (2006). The impact of presumed consent legislation on cadaveric organ donation: A cross country study. *Journal of Health Economics*, *25*, 599–620.

Becker, G. S., and Elias, J. J. (2007). Introducing incentives in the market for live and cadaveric organ donations. *Journal of Economic Perspectives*, *21*(3), 3–24.

Bellman, S., Johnson, E. J., Kobrin, S. J., and Lohse, G. L. (2004). International differences in information privacy concerns: A global survey of consumers. *Information Society*, *20*(5), 313–324.

Bellman, S., Johnson, E. J., and Lohse, G. L. (2001). To opt-in or opt-out? It depends on the question. *Communications of the ACM*, *44*(2), 25–27.

Benartzi, S., Peleg, E., and Thaler, R. H. (2009, January). Choice architecture and retirement saving plans. Paper presented at IMBS Colloquium. Institute for Mathematical Behavioral Sciences, University of California–Irvine.

Brown, C. L., and Krishna, A. (2004). The skeptical shopper: A metacognitive account for the effects of default options on choice. *Journal of Consumer Research*, *31*(3), 529.

Center for Democracy and Technology (2009, November 10). Refocusing the FTC's role in privacy protection. Policy Post. Washington, DC: Center for Democracy and Technology. Retrieved from https://www.cdt.org/policy/refocusing-ftc%E2%80%99s-role-privacy-protection

Choi, J., Laibson, D., Madrian, B., and Metric, A. (2001). *For better or for worse: Default effects and 401(k) savings behavior.* NBER Working Paper No. w8651. National Bureau of Economic Research.

Dinner, I., Johnson, E. J., Goldstein, D. G., and Kaiya, L. (2011). Partitioning default effects: Why people choose not to choose. *Journal of Experimental Psychology*, *17*(4), 332–341.

Fischhoff, B. (1991). Value elicitation: Is there anything in there? *American Psychologist*, *46*(8), 835–847.

Gabaix, X., Laibson, D., Moloche, G., and Weinberg, S. (2006). Costly information acquisition: Experimental analysis of a boundedly rational model. *American Economic Review*, *96*(4), 1043–1068.

Gäbel, H. (2002). *Donor and non-donor registries in Europe*. A report prepared on behalf of the committee of experts on the Organizational Aspects of Co-operation in Organ Transplantation of the Council of Europe. Stockholm, Sweden.

Gimbel, R. W., Strosberg, M. A., Lehrman, S. E., Gefenas, E., and Taft, F. (2003). Presumed consent and other predictors of cadaveric organ donation in Europe. *Progress in Transplantation*, *13*(1), 17–23.

Goldstein, D. G., Johnson, E. J., Herrmann, A., and Heitmann, M. (2008, December). Nudge your customers toward better choices. *Harvard Business Review*, pp. 100–105.

Henrich, J., Albers, W., Boyd, R., Gigerenzer, G., McCabe, K., Ockenfels, A., and Young, H. P. (2001). What is the role of culture in bounded rationality? In G. Gigerenzer and R. Selten (Eds.), *Bounded rationality: The adaptive toolbox* (pp. 343–360). Cambridge, MA: MIT Press.

Johnson, E. J., and Goldstein, D. (2003). Do defaults save lives? *Science*, *302*(5649), 1338–1339.

Johnson, E. J., Hershey, J., Meszaros, J., and Kunreuther, H. (1993). Framing, probability distortions, and insurance decisions. *Journal of Risk and Uncertainty*, *7*, 35–51.

Kerr, J. C. (2007, September 21). Numbers on Do Not Call list will start expiring next year. *USA Today*. Retrieved from http://www.usatoday.com/money/industries/telecom/2007-09-21-do-not-call_N.htm

Klassen, A. C., and Klassen, D. K. (1996). Who are the donors in organ donation? The family's perspective in mandated choice. *Annals of Internal Medicine*, *125*, 70–73.

Köszegi, B., and Rabin, M. (2006). A model of reference-dependent rreferences. *Quarterly Journal of Economics*, *121*(4), 1133–1165.

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics*, *112*(2), 443–477.

Madrian, B. C., and Shea, D. (2001). The power of suggestion: An analysis of 401(k) participation and saving behavior. *Quarterly Journal of Economics*, *116*(4), 1149–1187.

McKenzie, C.R.M., Liersch, M. J., and Finkelstein, S. R. (2006). Recommendations implicit in policy defaults. *Psychological Science*, *17*(5), 414–420.

O'Donoghue, T., and Rabin, M. (2001). Choice and procrastination. *Quarterly Journal of Economics*, *116*, 121–160.

Oz, M. C., Kherani, A. R., Rowe, A., Roels, L., Crandall, C., Tournatis, L., and Young, J. B. (2003). How to improve organ donation: Results of the ISHLT/FACT poll. *Journal of Heart and Lung Transplantation*, *22*(4), 389–396.

Payne, J. W., Bettman, J. R., and Johnson, E. J. (1992). Behavioral decision research: A constructive processing perspective. *Annual Review of Psychology*, *43*, 87–131.

Payne, J. W., Bettman, J. R., and Schkade, D. A. (1999). Measuring constructed preferences: Towards a building code. *Journal of Risk and Uncertainty*, *19*(1–3), 243–270.

Samuelson, W., and Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, *1*(1), 7–59.

Slovic, P. (1995). The construction of preference. *American Psychologist*, *50*(5), 364–371.

Sunstein, C. R., and Thaler, R. H. (2003). Libertarian paternalism is not an oxymoron. *University of Chicago Law Review*, *70*(4), 1159–1202.

Tversky, A., and Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *Quarterly Journal of Economics*, *106*(4), 1039–1061.

Wright, P. (2002). Marketplace metacognition and social intelligence. *Journal of Consumer Research*, *28*(4), 677–682.

Zauberman, G. (2003). The intertemporal dynamics of consumer lock-in. *Journal of Marketing Research*, *30*, 405–441.

CHAPTER 25

Choice Architecture

RICHARD H. THALER

CASS R. SUNSTEIN

JOHN P. BALZ

Consider the following hypothetical example:

The director of food services for a large city school system runs a series of experiments that manipulate the way in which the food is displayed in cafeterias. Not surprisingly, she finds that what the children eat depends on such things as the order of the items. Foods displayed at the beginning or end of the line are more likely to be eaten than items in the middle, and foods at eye level are more likely to be consumed than those in less salient locations. The question is, What use should the director make of this newfound knowledge?

Here are a few options to consider:

1. Arrange the food to make the students best off, all things considered.
2. Choose the food order at random.
3. Try to arrange the food to get the kids to pick the same foods they would choose on their own.
4. Maximize the sales of the items from the suppliers that are willing to offer the largest bribes.
5. Maximize profits, period.

Option 1 has obvious appeal. Although there can be some controversies, few would argue with the premise that the kids would be better off eating more fruits and vegetables and fewer burgers, fries, and sweets. Yes, this option might seem a bit intrusive, even paternalistic, but the alternatives are worse! Option 2, arranging the food at random, could be considered fair-minded and principled, and it is in one sense neutral. But from the perspective of a practical food service director, does it make any sense to scatter the ingredients to a salad bar at random through the line or separate the hamburgers from the buns? Also, if the orders are randomized across schools, then the children at some schools will have less healthy diets than those at other schools. Is this desirable?

Option 3 might seem to be an honorable attempt to avoid intrusion: try to mimic what the children would choose for themselves. Maybe this should be thought of as the objectively neutral choice, and maybe the director should neutrally follow people's wishes (at least where she is dealing with older students). But a little thought reveals that this is a difficult option to implement. The experiments prove that what kids choose depends on the order in which the items are displayed. What, then, are the true preferences of the children? What does it mean to try to devise a procedure for determining what the students would choose "on their own"? In a cafeteria, it is impossible to avoid some way of organizing food.

Option 4 might appeal to a corrupt cafeteria manager, and manipulating the order of the food items would put yet another weapon in the arsenal of available methods to exploit power. But if the director is honorable and honest this would not have any appeal. Like Options 2 and 3, Option 5 has some appeal, especially to a trained economist or a food-services director who is given incentives to follow this approach. But the school district must balance a range of priorities and requirements. Does it want its cafeterias to act as profit centers if the result is to make children less healthy?

In this example the director is what we call a *choice architect*. A choice architect has the responsibility for organizing the context in which people make decisions. Although this example is a figment of our imagination, many real people turn out to be choice architects, most without realizing it. Doctors describing the available treatments to patients, human-resource administrators creating and managing health-care plan enrollment, marketers devising sales strategies, ballot designers deciding where to put candidate names on a page, parents explaining the educational options available to a teenager; these are just a few examples of choice architects.

As the school cafeteria shows, small and apparently insignificant details can have major impacts on

people's behavior. A good rule of thumb is to assume that "everything matters." Even something as seemingly insignificant as the shape of a door handle. Early in Thaler's career, he taught a class on managerial decision making to business school students. Students would sometimes leave class early to go for job interviews (or a golf game) and would try to sneak out of the room as surreptitiously as possible. Unfortunately for them, the only way out of the room was through a large double door in the front in full view of the entire class (though not directly in Thaler's line of sight). The doors were equipped with large, handsome wood handles that were vertically mounted cylindrical pulls about two feet in length.

When the students came to these doors, they were faced with two competing instincts. One instinct says that to leave a room you push the door. This instinct is part of what psychologists call the reflective system, a deliberate and self-conscious thought process by which humans use logic and reasoning to help them make decisions. The other instinct says, when faced with large wooden handles that are obviously designed to be grabbed, you pull. This instinct is part of what is called the automatic system, a rapid, intuitive process that is not associated with what we would traditionally consider *thinking*.[1] It turns out that the latter instinct—the gut instinct—trumped the former—the conscious thought—and every student leaving the room began by pulling on the handle. Alas, the door opened outward.

At one point in the semester, Thaler pointed out this internal conflict to the class, as one embarrassed student was pulling on the door handle while trying to escape the classroom. Thereafter, as a student got up to leave, the rest of the class would eagerly wait to see whether the student would push or pull. Amazingly, most still pulled! Their automatic systems triumphed; the signal emitted by that big wooden handle simply could not be screened out.

Those doors are examples of poor architecture because they violate a simple psychological principle known as stimulus response compatibility, whereby the signal to be received (the stimulus) must be consistent with one's desired action. When signal and desire are in opposition, performance suffers and people blunder.

Consider, for example, the effect of a large, red, octagonal sign that reads "GO." The difficulties induced by such incompatibilities are easy to show experimentally. One of the most famous such demonstrations is the Stroop (1935) test. In the modern version of this experiment, people see words flashed on a computer screen and they have a very simple task. They press the right button if they see a word that is displayed in red, and press the left button if they see a word displayed in green. People find the task easy and can learn to do it very quickly with great accuracy. That is, until they are thrown a curve ball, in the form of the word *green* displayed in red, or the word *red* displayed in green. For these incompatible signals, response time slows and error rates increase. A key reason is that the automatic system reads the word faster than the color naming system can decide the color of the text. See the word *green* in red text and the nonthinking automatic system rushes to press the left button, which is, of course, the wrong one.

Although we have never seen a green stop sign, doors such as the ones described above are commonplace, and they violate the same principle. Flat plates say "push me" and big handles say "pull me," so do not expect people to push big handles! This is a failure of architecture to accommodate basic principles of human psychology. Life is full of products that suffer from such defects. Is it not obvious that the largest buttons on a television remote control should be the power, channel, and volume controls? Yet how many remotes have the volume control the same size as the "input" control button (which if pressed accidentally can cause the picture to disappear)?

This sort of design question is not a typical one for economists to think about because economists have a conception of human behavior that assumes, implicitly, that everyone relies completely on their reflective system, and a mighty good one at that! Economic agents are assumed to reason brilliantly, catalogue huge amounts of information that they can access instantly from their memories, and exercise extraordinary willpower. We call such creatures Econs. Plain old Humans make plenty of mistakes (even when they are consciously thinking!) and suffer all types of breakdowns in planning, self-control, and forecasting, as documented in many of the other chapters in this book.

Since the world is made up of Humans, not Econs, both objects and environments should be designed with Humans in mind. A great introduction to the topic of object design for humans is Donald Norman's wonderful book *The Design of Everyday Things* (1990). One of Norman's best examples is the design of a basic four-burner stove. Most such stoves have the burners in a symmetric arrangement, with the controls arranged in a linear fashion below. In this set-up, it is easy to get confused about which knob controls the front burner and which controls the back, and many pots and pans have been burned as a result.

Norman's basic lesson is that designers need to keep in mind that the users of their objects are Humans who are confronted every day with myriad choices and cues. The goal of this essay is to develop the same idea for people who create the environments in which we make decisions: *choice architects*. If you

indirectly influence the choices other people make, you have earned the title. Consider the person who designs the menu in a restaurant. The chef will have decided what food will be served, but it is someone else's job to put those offerings on paper (or blackboard), and there are lots of ways to do this. Should hot starters be in a different category from cold ones? Are pasta dishes a separate category? Within categories, how should dishes be listed? Where should prices be listed? In a world of Econs, these details would not matter, but for Humans, nearly everything matters, so choice architects can have considerable power to influence choices. Or to use our preferred language, they can nudge.

Of course, choice architects do not always have the best interests of the people they are influencing in mind. The menu designer may want to push profitable items or those about to spoil by printing them in bold print. Wily but malevolent nudgers, such as pushy mortgage brokers, can have devastating effects on the people who are influenced by them. Conscientious choice architects, however, do have the capability to self-consciously construct nudges in an attempt to move people in directions that will make their lives better. And since the choices these choice architects are influencing are going to be made by Humans, they will want their architecture to reflect a good understanding of how humans behave. In this chapter, we will offer some basic principles of effective choice architecture.

Defaults: Padding the Path of Least Resistance

For reasons of laziness, fear, and distraction, many people will take whatever option requires the least effort, or the path of least resistance. All these forces imply that if, for a given choice, there is a default option—an option that will obtain if the chooser does nothing—then we can expect a large number of people to end up with that option, whether or not it is good for them. These behavioral tendencies toward doing nothing will be reinforced if the default option comes with some implicit or explicit suggestion that it represents the normal or even the recommended course of action.

Defaults are ubiquitous and powerful. They are also unavoidable in the sense that for any node of a choice architecture system, there must be an associated rule that determines what happens to the decision maker if she does nothing. Of course, usually the answer is that if I do nothing, nothing changes; whatever is happening continues to happen. But not always. Some dangerous machines, such as chain saws and lawn mowers, are designed with "dead man switches," so that once a user lets go of the handle, the machine's blades stop. Some "big kid" slides at playgrounds are built with the first step about two feet off the ground to keep smaller kids from getting on and possibly hurting themselves.[2] When you leave a computer alone for a while to answer a phone call, nothing is likely to happen for a given period, after which the screen saver comes on. Neglect the computer long enough, and it may lock itself. Of course, a user can decide how long it takes before the screen saver comes on, but implementing that choice takes some action. Most computers come with a default time lag and a default screen saver. Chances are, those are the settings most people still have.

Downloading a new piece of software requires numerous choices, the first of which is "regular" or "custom" installation. Normally, one of the boxes is already checked, indicating it is the default. Which boxes do the software suppliers check? Two different motives are readily apparent: helpful and self-serving. Making the regular installation the default would be in the helpful category if most users will have trouble with the custom installation. Sending unwanted promotional spam to the user's email account would be in the self-serving category. In our experience, most software comes with helpful defaults regarding the type of installation, but many come with self-serving defaults on other choices. Just like choice architects, notice that not all defaults are selected to make the chooser's life easier or better.

Many organizations, public and private, have discovered the immense power of default options, big and small. Consider the idea of automatic renewal for magazine subscriptions. If renewal is automatic, many people will subscribe, for a long time, to magazines they do not read. Or the idea of automatically including seat reservations or travel insurance (for an extra charge, of course) when customers book train or airline tickets (Goldstein et al., 2008). Smart organizations have moved to double-sided printing as the default option. During the presidential campaign, Barack Obama's chief campaign advisor, David Plouffe, ordered all printers to be put on this setting, and the city of Tulsa, Oklahoma, estimates it will save more than $41,000 a year with double-sided printing (Simon, 2008).

The choice of the default can be quite controversial. Here are two examples. Faced with a budget crunch and the possible closing of some state parks because of the recent recession, Washington State legislators switched the default rule on state park fees that drivers pay when they renew their license plates. Before the recession, paying the $5 fee had been an option for drivers. The state switched from an opt-in to an opt-out arrangement, in which drivers are

charged unless they ask not to pay it. For transparency, the state provides information to each driver explaining the reason behind the change. So far, the move has worked, though critics do not think it is a long-term solution to the state's financial problems.

In another example, an obscure portion of the No Child Left Behind Act requires that school districts supply the names, addresses, and telephone numbers of students to the recruiting offices of the branches of the armed forces. However, the law stipulates that "a secondary school student or the parent of the student may request that the student's name, address, and telephone listing not be released without prior written parental consent, and the local educational agency or private school shall notify parents of the option to make a request and shall comply with any request" (NCLB, 2002). Some school districts, such as Fairport, New York, interpreted this law as allowing them to implement an opt-in policy. That is, parents were notified that they could elect to make their children's contact information available, but if they did not do anything, this information would be withheld. This reading of the law did not meet with the approval of then-Secretary of Defense Donald Rumsfeld. The Departments of Defense and Education sent a letter to school districts asserting that the law required an opt-out implementation. Only if parents actively requested that the contact information on their children be withheld would that option apply. In typical bureaucratic language, the departments contended that the relevant laws "do not permit LEA's [local educational agencies] to institute a policy of not providing the required information unless a parent has affirmatively agreed to provide the information."[3] Both the Department of Defense and the school districts realized that opt-in and opt-out policies would lead to very different outcomes. Not surprisingly, much hue and cry ensued.

We have emphasized that default rules are inevitable—that private institutions and the legal system cannot avoid choosing them. In some cases, though not all, there is an important qualification to this claim. The choice architect can force the choosers to make their own choice. We call this approach *required choice*, or *mandated choice*. In the software example, required choice would be implemented by leaving all the boxes unchecked and by requiring that at every opportunity one of the boxes be checked in order for people to proceed. In the case of the provision of contact information to the military recruiters, one could imagine a system in which all students (or their parents) are required to fill out a form indicating whether they want to make their contact information available. For emotionally charged issues like this one, such a policy has considerable appeal, because people might not want to be defaulted into an option that they might hate (but fail to reject because of inertia or real, or apparent, social pressure).

A good example where mandated choice has considerable appeal is organ donation. As discussed by Johnson and Goldstein (2003) some countries have adopted an opt-out approach to organ donation called *presumed consent*. This approach clearly maximizes the number of people who (implicitly) agree to make their organs available. However, some people strenuously object to this policy, feeling that the government should not presume anything about their organs. An effective compromise is mandated choice. For example, in Illinois when drivers go to get their license renewed and a new photograph taken they are required to answer the question, Do you wish to be an organ donor? before they can get their license. This policy has produced a 60% sign-up rate compared to the national average of 38%.[4] Furthermore, since the choice to be a donor is explicit rather than implicit, family members of deceased donors are less likely to object.

We believe that required choice, which is favored by many who like freedom, is sometimes the best way to go. But consider two points about the approach. First, Humans will often consider required choice to be a nuisance or worse and would much prefer to have a good default. In the software example, it is helpful to know what the recommended settings are. Most users do not want to have to read an incomprehensible manual in order to determine which arcane setting to elect. When choice is complicated and difficult, people might greatly appreciate a sensible default. It is hardly clear that they should be forced to choose.

Second, required choosing is generally more appropriate for simple yes-or-no decisions than for more complex choices. At a restaurant, the default option is to take the dish as the chef usually prepares it, with the option to substitute or remove certain ingredients. In the extreme, required choosing would imply that the diner has to give the chef the recipe for every dish she orders! When choices are highly complex, required choosing may not be a good idea; it might not even be feasible.

Expect Error

Humans make mistakes. A well designed system expects its users to err and is as forgiving as possible. Some examples from the world of real design illustrate this point.

In the Paris subway system, Le Métro, users insert a paper card the size of a movie ticket

into a machine that reads the card, leaves a record on the card that renders it "used," and then spits it out from the top of the machine. The cards have a magnetic strip on one side but are otherwise symmetric. Intelligent subway card machines are able to read the strip no matter which way a user inserts her card. In stark contrast to Le Métro is the system used in most Chicago parking garages. When entering the garage, a driver puts a credit card into a machine that reads it and remembers the information. Then when leaving, the driver inserts the card again into another machine at the exit. This involves reaching out of the car window and inserting the card into a slot. Because credit cards are not symmetric, there are four possible ways to put the card into the slot (face up or down, strip on the right or left). Exactly one of those ways is the right way. And in spite of a diagram above the slot, it is very easy to put the card in the wrong way, and when the card is spit back out, it is not immediately obvious what caused the card to be rejected or to recall which way it was inserted the first time.

Over the years, automobiles have become much friendlier to their Human operators. They buzz when the seat belts are not buckled. Warning signs flash when the gas gauge is low, or the oil life is almost over. Many cars come with an automatic switch for the headlights that turns them on when the car is operating and off when it is not, eliminating the possibility of leaving lights on overnight and draining the battery.

But some error-forgiving innovations are surprisingly slow to be adopted. Take the case of the gas tank cap. On any sensible car the gas cap is attached by a piece of plastic, so that when a driver removes the cap she cannot drive off without it. This plastic cap is so inexpensive that once one firm had the good idea to include this feature, there should be no excuse for building a car without one.

Leaving the gas cap behind is a special kind of predictable error psychologists call a *postcompletion* error (Byrne and Bovair, 1997). The idea is that once the main task is finished, people tend to forget things relating to previous steps. Other examples include leaving ATM cards in the machine after withdrawing cash, or leaving the original in the copying machine after making copies. Most ATMs (but not all) no longer allow this error because the card is returned immediately. Another strategy, suggested by Norman, is to use what he calls a *forcing function*. In order to accomplish a desire, another step must first be taken. If a user has to remove her card before physically receiving her cash, she will not forget it.

Another automobile-related bit of good design involves the nozzles for different varieties of gasoline. The nozzles that deliver diesel fuel are too large to fit into the opening on cars that use gasoline, so it is not possible to make the mistake of putting diesel fuel in a gasoline-powered car (though it is still possible to make the opposite mistake). The same principle has been used to reduce the number of errors involving anesthesia. One study found that human error (rather than equipment failure) caused 82% of the "critical incidents." A common error was that the hose for one drug was hooked up to the wrong delivery port, so the patient received the wrong drug. This problem was solved by designing the equipment so that the gas nozzles and connectors were different for each drug. It became physically impossible to make this previously frequent mistake (Vicente, 2006).

A major problem in health care that costs billions of dollars annually is called *drug compliance*. Many patients, especially the elderly, are on medicines they must take regularly and in the correct dosage. So here is a choice-architecture question: How should a drug designer construct a dosage schedule?

If a onetime dose administered immediately by the doctor (which would be best on all dimensions but is often technically infeasible) is ruled out, then the next-best solution is a medicine taken once a day, preferably in the morning. It is clear why once a day is better than twice (or more) a day. Because the more often a patient must take the drug, the more opportunities she has to forget. But frequency is not the only concern; regularity is also important. Once a day is much better than once every other day because this schedule activates the automatic system. Taking the pill becomes a habit. By contrast, remembering to take medicine every other day is beyond most Humans. (Similarly, meetings that occur every week are easier to remember than those that occur every other week.) Some medicines are taken once a week, and most patients take this medicine on Sundays (because that day is different from other days for most people and thus easy to associate with taking one's medicine).

Birth control pills present a special problem along these lines, because they are taken every day for three weeks and then skipped for one week. To solve this problem and to make the process automatic, the pills

are typically sold in a special container that contains twenty-eight pills, each in a numbered compartment. Patients are instructed to take a pill every day, in order. The pills for days twenty-two through twenty-eight are placebos whose only role is to facilitate compliance for Human users.

Another serious problem in the world of medicine stems from the often frenzied hospital environment. Because a patient's medical care can require hundreds of decisions each day, some doctors and hospital administrators have experimented with using checklists for certain treatments where human error can lead to serious harm. The checklists contain simple, routine actions, all of which doctors learned in medical school but may simply forget to follow because of time constraints, stress, or distractions. For instance, the checklist designed by a critical-care specialist at Johns Hopkins Hospital for treating line infections included five simple steps from washing one's hands with soap to putting a sterile dressing over the catheter site once the line is in.

The point of the checklists was twofold. It helped with memory recall, which is critical in a hospital where events like a person writhing in pain can easily make you forget about whether you have washed your hands. The checklist also broke down the entire complex process into a series of steps that allowed staffers to better see what constituted a high standard of performance. The results from what seem like just simple reminders stunned the doctors. The ten-day line-infection rate fell from 11% to zero. After fifteen more months, only two patients got line infections. Forty three infections and eight deaths had been prevented. Two million dollars had been saved (Gawande, 2007, 2010; Pronovost et al., 2006).

While working on *Nudge* (Thaler and Sunstein, 2008), Thaler sent an email to Google's chief economist, Hal Varian. He intended to attach a draft of the introduction to give Varian an overview of the book but forgot the attachment. When Varian wrote back to ask for the missing attachment, he noted that Google was experimenting with a new feature on its email program, Gmail, that would solve this problem. A user who mentions the word attachment but does not include one would be prompted with "Did you forget your attachment?" Thaler sent the attachment along and told Varian that this was exactly what the book was about.

Visitors to London who come from the United States or Europe have a problem being safe pedestrians. They have spent their entire lives expecting cars to come at them from the left, and their automatic system knows to look that way. But in the United Kingdom automobiles drive on the left-hand side of the road, and so the danger often comes from the right. Many pedestrian accidents occur as a result. The city of London tries to help with good design. On many corners, especially in neighborhoods frequented by tourists, the pavement has signs that say, "Look right!"

Give Feedback

The best way to help Humans improve their performance is to provide feedback. Well-designed systems tell people when they are doing well and when they are making mistakes. Some examples are the following:

Digital cameras generally provide better feedback to their users than film cameras. After each shot, the photographer can see a (small) version of the image just captured. This eliminates errors that were common in the film era, from failing to load the film properly (or at all), to forgetting to remove the lens cap, to cutting off the head of the central figure of the picture. However, early digital cameras failed on one crucial feedback dimension. When a picture was taken, there was no audible cue to indicate that the image had been captured. Modern models now include a satisfying, but completely fake, shutter click sound when a picture has been taken. Some cell phones, especially those aimed at the elderly, include a fake dial tone, for similar reasons.

One of the most scenic urban highways in the world is Chicago's Lake Shore Drive, which hugs the Lake Michigan coastline that is the city's eastern boundary. The drive offers stunning views of Chicago's magnificent skyline. There is one stretch of this road that puts drivers through a series of S curves. These curves are dangerous. Many drivers fail to take heed of the reduced speed limit (25 mph) and wipe out. In September 2006, the city adopted a new strategy for slowing traffic. It painted a series of white lines perpendicular to the traveling cars. The lines progressively narrow as drivers approach the sharpest point of the curve, giving them the illusion of speeding up, and nudging them to tap their brakes.

Until the recent release of data by the Chicago Department of Transportation, only anecdotal accounts provided any indication of how effective the lines had been in preventing accidents. According to

an analysis conducted by city traffic engineers, there were 36% fewer crashes in the six months after the lines were painted compared to the same six-month period the year before (September 2006–March 2007 and September 2005–March 2006). This level of reduction at the cost of some extra paint is remarkable. To see if it could make the road even safer, the city installed a series of overhead flashing beacons, yellow and black chevron alignment signs, and warning signs posting the reduced advisory speed limit. Again, accidents fell—47% over a six-month period (March 2007–August 2007 and March 2006–August 2006). Keep in mind that this post-six-month-period effect included both the signs and the lines. The city considers both numbers to be signs of success.

An important type of feedback is a warning that things are going wrong, or, even more helpful, are about to go wrong. Laptops warn users to plug in or shut down when the battery is dangerously low. But warning systems have to avoid the "boy who cried wolf" problem of offering so many warnings that they are ignored. If a computer constantly nags users about whether they want to open attachments, they begin to click "yes" without thinking about it. These warnings are thus rendered useless.

Some clever feedback systems are popping up in ways that are good for the environment and household budgets. There is the Ambient Orb, a small ball that glows red when a customer is using lots of energy but green when energy use is modest. Utility companies have experimented with sending customers electricity bills that tell them how much energy they are using compared to their neighbors. Prius drivers already know how easy it is to be entranced by a screen that continuously updates your miles-per-gallon rate, and how hard it can be not to adjust driving in order to squeeze the most mileage out of each fuel tank. Nissan has developed an acceleration pedal that adjusts its resistance when the driver has a lead foot (NASCAR-like acceleration wastes gas). Two Stanford graduate students have come up with a piece of technology that combines all of these feedback mechanisms into one amazing piece of choice architecture. Called the SmartSwitch, users turn a light on using a slide switch. Like Nissan's pedal, the switch is harder to push when lots of energy is being used, giving the owner a subtle reminder about those bad habits. The switch can also be linked to other homeowners in the neighborhood so that the switch slides less smoothly when all the neighbors are blasting their air conditioners on a hot day.

Feedback can be improved in many activities. Consider the simple task of painting a ceiling. This task is more difficult than it might seem because ceilings are nearly always painted white, and it can be hard to see exactly where you have painted. Later, when the paint dries, the patches of old paint will be annoyingly visible. How to solve this problem? Some helpful person invented a type of ceiling paint that goes on pink when wet but turns white when dry. Unless the painter is so color-blind that he cannot tell the difference between pink and white, this solves the problem.

Understanding Mappings: From Choice to Welfare

Some tasks are easy, like choosing a flavor of ice cream; other tasks are hard, like choosing a medical treatment. Consider, for example, an ice cream shop where the varieties differ only in flavor, not calories or other nutritional content. Selecting which ice cream to eat is merely a matter of choosing the one that tastes best. If the flavors are all familiar, such as vanilla, chocolate, and strawberry, most people will be able to predict with considerable accuracy the relation between their choice and their ultimate consumption experience. Call this relation between choice and welfare a *mapping*. Even if there are some exotic flavors, the ice cream store can solve the mapping problem by offering a free taste.

Choosing among treatments for some disease is quite another matter. Suppose a person is diagnosed with prostate cancer and must choose among three options: surgery, radiation, and watchful waiting (which means do nothing for now). Each of these options comes with a complex set of possible outcomes regarding side effects of treatment, quality of life, length of life, and so forth. Comparing the options involves making trade-offs between a longer life and an increased risk of unpleasant side effects, such as impotence or incontinence. Weighing these scenarios makes for a hard decision at two levels. The patient is unlikely to know these trade-offs, and he is unlikely to be able to imagine what life would be like if he were incontinent. Yet here are two scary facts about this scenario. First, most patients decide which course of action to take in the very meeting at which their doctor breaks the bad news about the diagnosis. Second, the treatment option they choose depends strongly on the type of doctor they see (Zeliadt et al., 2006). (Some specialize in surgery, others in radiation. None specialize in watchful waiting. Guess which option is the most likely candidate for underutilization?)

The comparison between ice cream and treatment options illustrates the concept of mapping. A good system of choice architecture helps people improve

their ability to map and hence to select options that will make them better off. One way to do this is to make the information about various options more comprehensible by transforming numerical information into units that translate more readily into actual use. When buying apples to make into apple cider, it helps to know the rule of thumb that it takes three apples to make one glass of cider.

Mapping is a frequent problem in consumer electronic decisions like purchasing a digital camera. Cameras advertise their megapixels, and the impression created is certainly that the more megapixels the better. This assumption is itself subject to question, because photos taken with more megapixels take up more room on the camera's storage device and a computer's hard drive. But what is most problematic for consumers is translating megapixels (not the most intuitive concept) into understandable terms that help them order their preferences. Is it worth paying an additional hundred dollars to go from four to five megapixels? Suppose instead that manufacturers listed the largest print size recommended for a given camera. Instead of being given the options of three, five, or seven megapixels, consumers might be told that the camera can produce quality photos at 4 by 6 inches, 9 by 12, or poster size.

Often people have a problem in mapping products into money. For simple choices, of course, such mappings are trivial. If a Snickers bar costs $1, it is easy to figure out the cost of a Snickers bar every day. But do consumers know how much it costs you to use a credit card? Among the many built-in fees are (1) an annual fee for the privilege of using the card (common for cards that provide benefits such as frequent-flier miles); (2) an interest rate for borrowing money (which depends on your deemed credit worthiness); (3) a fee for making a payment late (and you may end up making more late payments than you anticipate); (4) interest on purchases made during the month that is normally not charged if your balance is paid off but begins if you make your payment one day late; (5) a charge for buying things in currencies other than dollars; and (6) the indirect fee of higher prices that retailers pass along to consumers to offset the small percentage of each transaction the credit card companies take.

Credit cards are not alone in having complex pricing schemes that are neither transparent nor comprehensible to consumers. Think about mortgages, cell phone calling plans, and auto insurance policies, just to name a few. For these and related domains, we propose a very mild form of government regulation that we call RECAP: Record, Evaluate, and Compare Alternative Prices.

Here is how RECAP would work in the cell phone market. The government would not regulate how much issuers could charge for services, but it would regulate their disclosure practices. The central goal would be to inform customers of every kind of fee that currently exists. This would not be done by printing a long unintelligible document in fine print. Instead, issuers would be required to make public their fee schedule in a spreadsheet-like format that would include all relevant formulas. Suppose an American is visiting Toronto and his cell phone rings. How much is it going to cost to answer it? What if he downloads some email? All these prices would be embedded in the formulas. This is the price disclosure part of the regulation.

The usage disclosure requirement would be that once a year, issuers would have to send their customers a complete listing of all the ways they had used the phone and all the fees that had been incurred. This report would be sent two ways, by mail and, more important, electronically. The electronic version would also be stored and downloadable on a secure website.

Producing the RECAP reports would cost cell phone carriers very little, but the reports would be extremely useful for customers who want to compare the pricing plans of cell phone providers, especially after they had received their first annual statement. Private websites similar to existing airline and hotel sites would emerge to allow an easy way to compare services. With just a few quick clicks, a shopper would easily be able to import her usage data from the past year and find out how much various carriers would have charged, given her usage patterns.[5] Consumers who are new to the product (getting a cell phone for the first time, for example) would have to guess usage information for various categories, but the following year they could take full advantage of the system's capabilities. Already sites like this are popping up. One of them, billshrink.com, tracks cell phone plans, credit cards, and gas stations, saving people money by helping them pick the best plan (or card) for their consumer habits. We think that in many other domains, from mortgages to energy use to Medicare, a RECAP program could greatly improve people's ability to make good choices.

Structure Complex Choices

People adopt different strategies for making choices depending on the size and complexity of the available options. When facing a small number of well-understood alternatives, the tendency is to examine all the attributes of all the alternatives and then make trade-offs when necessary. But when the choice set gets large, alternative strategies must be employed, leading to serious problems.

Consider, for example, someone who has just been offered a job at a company located in another

city. Compare two choices: which office to select and which apartment to rent. Suppose this individual is offered a choice of three available workplace offices. A reasonable strategy is to look at all three offices, note the ways they differ, and then make some decisions about the importance of such attributes as size, view, neighbors, and distance to the nearest restroom. This is described in the choice literature as a *compensatory* strategy, since a high value for one attribute (big office) can compensate for a low value for another (loud neighbor).

Obviously, the same strategy cannot be used to pick an apartment. In any large city, thousands of apartments are available, and no single person can see them all. Instead, the task must be simplified. One strategy to use is what Tversky (1972) called *elimination by aspects*. Someone using this strategy first decides what aspect is most important (say, commuting distance), establishes a cutoff level (say, no more than a thirty minute commute), and then eliminates all alternatives that do not meet this standard. The process is repeated, attribute by attribute until either a choice is made or the set is narrowed down enough to switch over to a compensatory evaluation of the "finalists."

When people are using a simplifying strategy of this kind, alternatives that do not meet the minimum cutoff scores may be eliminated even if they are high on all other dimensions. For example, an apartment with a thirty-five minute commute will not be considered even if it has an ocean view and costs $200 a month less than any of the alternatives.

Social science research reveals that as the choices become more numerous or vary on more dimensions or both, people are more likely to adopt simplifying strategies. The implications for choice architecture are related. As alternatives become more numerous and more complex, choice architects have more to think about and more work to do and are much more likely to influence choices (for better or for worse). For an ice cream shop with three flavors, any menu listing those flavors in any order will do just fine, and effects on choices (such as order effects) are likely to be minor because people know what they like. As the choices become more numerous, though, good choice architecture will provide structure, and structure will affect outcomes.

Consider the example of a paint store. Even ignoring the possibility of special orders, paint companies sell more than two thousand colors for a home's walls. It is possible to think of many ways of structuring how those paint colors are offered to the customer. Imagine, for example, that the paint colors were listed alphabetically. Arctic White might be followed by Azure Blue, and so forth. While alphabetical order is a satisfactory way to organize a dictionary (at least if you have a guess as to how a word is spelled), it is a lousy way to organize a paint store.

Instead, paint stores have long used something like a paint wheel, with color samples ordered by their derivation from the three primary colors: all the blues are together, next to the greens, and the reds are located near the oranges, and so forth. The problem of selection is made considerably easier by the fact that people can see the actual colors, especially because the names of the paints are typically uninformative. (On the Benjamin Moore Paints website, three similar shades of beige are called "Roasted Sesame Seed," "Oklahoma Wheat," and "Kansas Grain.")

Thanks to modern computer technology and the World Wide Web, many problems of consumer choice have been made simpler. The Benjamin Moore Paints website not only allows the consumer to browse through dozens of shades of beige, but it also permits the consumer to see (within the limitations of the computer monitor) how a particular shade will work on the walls with the ceiling painted in a complementary color. And the variety of paint colors is small compared to the number of books sold by Amazon (millions) or web pages covered by Google (billions). Many companies such as Netflix, the mail-order DVD rental company, succeed in part because of immensely helpful choice architecture. Customers looking for a movie to rent can easily search movies by actor, director, genre, and more, and if they rate the movies they have watched, they can also get recommendations based on the preferences of other movie lovers with similar tastes, a method called *collaborative filtering*. People use the judgments of other people who share their tastes to filter through the vast number of books or movies available in order to increase the likelihood of picking one they like. Collaborative filtering is an effort to solve a problem of choice architecture. If an individual knows what others like him tend to like, he might be comfortable in selecting unfamiliar products. For many, collaborative filtering saves cognitive resources and search costs, thus making difficult choices easier.

A cautionary note: surprise and serendipity can be fun—and salutary, too—and there may be disadvantages if the primary source of information is what people like us like. Sometimes it is good to learn what people unlike us like and test it out. For fans of the mystery writer Robert B. Parker, collaborative filtering will probably direct them to other mystery writers, not Joyce Carol Oates or Henry James. Perhaps second-generation collaborative filtering will also present users with potential surprises. Democrats who like books that fit their predilections might want to see what Republicans are arguing because no party can possibly have a monopoly on wisdom.

Public-spirited choice architects—those who run the daily newspaper, for example—know that it is good to nudge people in directions that they might not have specifically chosen in advance. Structuring choice sometimes means helping people to learn so they can later make better choices on their own.[6]

Incentives

Our last topic is the one with which most economists would have started: prices and incentives. Although we have been stressing factors that are often neglected by traditional economic theory, we do not intend to suggest that standard economic forces are unimportant. This is as good a point as any to state for the record that we believe in supply and demand. If the price of a product goes up, suppliers will usually produce more of it and consumers will usually want less of it. So choice architects must think about incentives when they design a system. Sensible architects will put the right incentives on the right people. One way to start to think about incentives is to ask four questions about a particular choice architecture:

Who uses?

Who chooses?

Who pays?

Who profits?

Free markets often solve the key problems of decision making by giving people an incentive to make good products and to sell them at the right price. If the market for sneakers is working well, abundant competition will drive bad sneakers (meaning those that do not provide good value to consumers at their price point) from the marketplace, and price the good ones in accordance with people's tastes. Sneaker producers and sneaker purchasers have the right incentives. But sometimes incentive conflicts arise. Consider a simple case. Two friends go for a weekly lunch and each chooses his own meal and pays for what he eats. The restaurant serves their food and keeps their money. No conflicts here. Now suppose they decide to take turns paying for each other's lunch. Each now has an incentive to order something more expensive on the weeks that the other is paying, and vice versa. (In this case, though, friendship introduces a complication; good friends may well order something cheaper if he knows that the other is paying. Sentimental but true.)

Many markets (and choice architecture systems) are replete with incentive conflicts. Perhaps the most notorious is the U.S. health-care system. The patient receives the health-care services that are chosen by his physician and paid for by the insurance company, with intermediaries from equipment manufacturers to drug companies to malpractice lawyers extracting part of the original cost. Different intermediaries have different incentives, and the results may not be ideal for either patients or doctors. Of course, this point is obvious to anyone who thinks about these problems. But as usual, it is possible to elaborate and enrich the standard analysis by remembering that the agents in the economy are Humans. To be sure, even mindless Humans demand less when they notice that the price has gone up, but only if they are paying enough attention to notice the change in price.

The most important modification that must be made to a standard analysis of incentives is salience. Are choosers aware of the incentives they face? In free markets, the answer is usually yes, but in important cases, the answer is no. Consider the example of members of an urban family deciding whether to buy a car. Suppose their choices are to take taxis and public transportation or to spend $10,000 to buy a used car, which they can park on the street in front of their home. The only salient costs of owning this car will be the stops at the gas station, occasional repair bills, and a yearly insurance bill. The opportunity cost of the $10,000 is likely to be neglected. (In other words, once they purchase the car, they tend to forget about the $10,000 and stop treating it as money that could have been spent on something else.) In contrast, every time the family uses a taxi, the cost will be in their face, with the meter clicking every few blocks. So a behavioral analysis of the incentives of car ownership will predict that people will underweight the opportunity costs of car ownership, and possibly other less salient aspects such as depreciation, and may overweight the very salient costs of using a taxi.[7] An analysis of choice architecture systems must make similar adjustments.

Of course, salience can be manipulated, and good choice architects can take steps to direct people's attention to incentives. The telephones at the INSEAD School of Business in France are programmed to display the running costs of long distance phone calls. To protect the environment and increase energy independence, similar strategies could be used to make costs more salient in the United States. Suppose home thermostats were programmed to announce the cost per hour of lowering the temperature a few degrees during the heat wave. This would probably have more effect on behavior than quietly raising the price of electricity, a change that will be experienced only at the end of the month when the bill comes. Suppose in this light that government wants to increase energy conservation. Increases in the price of electricity will surely have an effect; making the increases salient

will have a greater effect. Cost-disclosing thermostats might have a greater impact than (modest) price increases designed to decrease use of electricity. Google, for instance, has developed a free electricity usage monitoring tool that provides information on energy usage and that, for customers without smart thermostats, can be hooked up to a handheld device.

In some domains, people may want the salience of gains and losses treated asymmetrically. For example, no one would want to go to a health club that charged its users on a per-step basis on the Stairmaster. However, many Stairmaster users enjoy watching the "calories burned" meter while they work out (especially since those meters seem to give generous estimates of calories actually burned). In Japan, some treadmills display pictures of food, like coffee and ice cream, during the workout to allow users to better balance their exercise and dieting habits.

We have sketched six principles of good choice architecture. As a concession to the bounded memory of our readers, we thought it might be useful to offer a mnemonic device to help recall the six principles. By rearranging the order, and using one small fudge, the following emerges.

iNcentives

Understand mappings

Defaults

Give feedback

Expect error

Structure complex choices

Voilà: NUDGES

With an eye on these nudges, choice architects can improve the outcomes for their Human users.

Notes

This essay draws heavily on Thaler and Sunstein's book *Nudge* (2008) and on other material that has appeared on the book's blog (www.nudges.org), and was edited by Balz. This chapter was written well before Sunstein joined the Obama Administration as counselor to the director of the Office of Management and Budget, later to be confirmed as administrator of the Office of Information and Regulatory Affairs. Nothing said here represents an official position in any way. Thaler is a professor at the Booth School of Business, University of Chicago. Sunstein is a professor at the Harvard Law School. Balz received his Ph.D. in political science from the University of Chicago.

1. In the psychology literature, these two systems are sometimes referred to as System 2 and System 1, respectively.

2. Thanks to a *Nudge* reader for this example.

3. Letter of July 2, 2003, to State School Officers signed by William Hanse, deputy secretary of education, and David Chu, undersecretary of defense.

4. Illinois's organ donation rate is compiled by Donate Life Illinois (http://www.donatelifeillinois.org/). For the national organ donor rate see Donate Life America (2009).

5. We are aware, of course, that behavior depends on prices. If my current cell phone provider charges me a lot to make calls in Canada and I react by not making such calls, I will not be able to judge the full value of an alternative plan with cheap calling in Canada. But where past usage is a good predictor of future usage, a RECAP plan would be very helpful.

6. Sunstein (2007), explores this point in detail.

7. Companies such as Zipcar that specialize in short-term rentals could profitably benefit by helping people solve these mental accounting problems.

References

Byrne, M. D., and Bovair, S. (1997). A working memory model of a common procedural error. *Cognitive Science*, *21*, 31–61.

City of Tulsa (Oklahoma). (2009). *City hall's new printing policies expected to reduce costs.* Retrieved from http://www.cityoftulsa.org/COTLegacy/Enews/2009/3-3/SAVINGS.ASP

Donate Life America. (2009). *National donor designation report card*. Retrieved from http://www.donatelife.net/donante/DLA+Report+Card+2009.pdf

Gawande, A. (2007, December 10). The checklist. *New Yorker*, pp. 86–95.

———. (2010). *The checklist manifesto: How to get things right*. New York: Metropolitan Books.

Goldstein, D. G., Johnson, E. J., Herrmann, A., and Heitmann, M. (2008, December). Nudge your customers toward better choices. *Harvard Business Review*, pp. 99–105.

Johnson, E. J., and Goldstein, D. (2003). Do defaults save lives? *Science*, *302*(5649), 1338–1339.

No Child Left Behind (NCLB) Act of 2001. Pub. L. 107–110, 115 State. 1425 (2002).

Norman, D. (1990). *The design of everyday things*. Sydney: Currency.

Pronovost, P., Needham, D., Berenholtz, S., Sinopoli, D., Chu, H., Cosgrove, S., et al. (2006). An intervention to decrease catheter-related bloodstream infections in the ICU. *New England Journal of Medicine*, *355*(26), 2725–2732.

Simon, R. (2008). *Relentless: How Barack Obama outsmarted Hillary Clinton*. Retrieved from http://www.politico.com/relentless/

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *12*, 643–662.

Sunstein, C. R. (2007). *Republic.com 2.0*. Princeton, NJ: Princeton University Press.

Sunstein, C. R., and Thaler, R. H. (2003). Libertarian paternalism is not an oxymoron. *University of Chicago Law Review*, *70*, 1159–1202.

Thaler, R. H., and Sunstein, C. R. (2003). Libertarian paternalism. *American Economic Review*, *93*(2), 175–179.

———. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. New Haven, CT: Yale University Press.

Thaler, R. H., Sunstein, C. R., and Balz, J. P. (2010). *Choice architecture*. Unpublished manuscript. Retrieved from http://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID1583509_code1460474.pdf?abstractid=1583509&mirid=1

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, *76*, 31–48.

Van De Veer, D. (1986). *Paternalistic intervention: The moral bounds on benevolence*. Princeton, NJ: Princeton University Press.

Vicente, K. J. (2006). *The human factor: Revolutionizing the way people live with technology*. New York: Routledge.

Zeliadt, S. B., Ramsey, S. D., Penson, D. F., Hall, I. J., Ekwueme, D. U., Stroud, L., and Lee, J. W. (2006). Why do men choose one treatment over another? *Cancer*, *106*, 1865–1874.

CHAPTER 26

Behaviorally Informed Regulation

MICHAEL S. BARR

SENDHIL MULLAINATHAN

ELDAR SHAFIR

Policy makers typically approach human behavior from the perspective of the *rational agent* model, which relies on normative, a priori analyses. The model assumes people make insightful, well-planned, highly controlled, and calculated decisions guided by considerations of personal utility. This perspective is promoted in the social sciences and in professional schools and has come to dominate much of the formulation and conduct of policy. An alternative view, developed mostly through empirical behavioral research, and the one we will articulate here, provides a substantially different perspective on individual behavior and its policy and regulatory implications. According to the empirical perspective, behavior is the amalgam of perceptions, impulses, judgments, and decision processes that emerge from the impressive machinery that people carry behind the eyes and between the ears. Actual human behavior, it is argued, is often unforeseen and misunderstood by classical policy thinking. A more nuanced behavioral perspective, it is suggested, can yield deeper understanding and improved regulatory insight.

For a motivating example, consider the recent mortgage crisis in the United States. While the potential causes are myriad, a central problem was that many borrowers were offered and took out loans that they did not understand and could not afford, with disastrous results for the borrowers, financial firms, and the national economy. Borrowers, like most people, are not well described by the rational agent model. At the same time, we argue, a behavioral perspective that focuses only on the individual is incomplete for policy purposes. In some contexts, firms have strong incentives to exploit consumer biases and will shape their conduct in response not only to the behavior of consumers but also to the actions of regulators. Thus, policy also needs to take into account market contexts and the incentives and behaviors that they afford firms.

In what follows, we will outline some of the main research underpinning the behavioral perspective pertinent to regulation. We will explore how firms interact with consumers in different market contexts and will propose a model for understanding this interaction. We will then develop an analytic framework for behaviorally informed regulation and conclude with examples of relevant policy applications.

On Behavior

In contrast with the classical theory, which is driven by rational agents who make well-informed, carefully considered, and fully controlled choices, behavioral research has shown that individuals depart from this decision-making model in important ways. Among other things, the availability and dissemination of data do not always lead to effective communication and knowledge; understanding and intention do not necessarily lead to the desired action; and purportedly inconsequential contextual nuances, whether intentional or not, can shape behavior and alter choices, often in ways that people themselves agree diminish their well-being in unintended ways. Individuals often exhibit temporal biases and misforecast their own behavior. By way of illustration, we will highlight how context, decisional conflict, mental accounting, knowledge and attention constraints, and institutions, shape individual decisions and behavior.

Context

Human behavior turns out to be heavily context dependent, a function of both the person and the situation. One of the major lessons of modern psychological research is the impressive power that the situation exerts, along with a persistent tendency to underestimate that power relative to the presumed influence of intention, education, or personality traits. In his now-classic obedience studies, for example, Milgram (1974) showed how decidedly mild situational pres-

sures sufficed to generate persistent willingness, against their own wishes, on the part of individuals to administer what they believed to be grave levels of electric shock to innocent subjects. Along similar lines, Darley and Batson (1973) recruited seminary students to deliver a practice sermon on the parable of the Good Samaritan. While half the seminarians were told they had plenty of time, others were led to believe they were running late. On their way to give the talk, all participants passed an ostensibly injured man slumped in a doorway groaning. Whereas the majority of those with time to spare stopped to help, a mere 10% of those who were running late stopped, while the remaining 90% stepped over the victim and rushed along. In contrast with these people's ethical training, scholarship, and presumed character, the contextual nuance of a minor time constraint proved decisive in the decision of whether to stop to help a suffering man. The powerful impact of context on behavior, we argue, increases the importance of effective regulation and regulators' responsibility to assess effectiveness in policy contexts.

Context is made all the more important because an individual's predictions about her behavior in the future are often made in contexts different from those in which the individual will later find herself. Koehler and Poon (2005; See Lewin, 1951) argued that people's predictions of their future behavior overweight the strength of their current intentions and underweight the contextual factors that influence the likelihood that those intentions will translate into action. This imbalance can generate systematically misguided plans among consumers, who, reassured by their good intentions, proceed to put themselves in ill-conceived situations that are powerful enough to make them act and choose otherwise.

Decisional Conflict

Three decades of behavioral research have led to the notion that people's preferences are typically constructed, not merely revealed, during the decision making process (Lichtenstein and Slovic, 2006). The construction of preferences is heavily influenced by the nature and the context of decision. For example, the classical view of decision making does not anticipate that decisional conflict will influence the making of decisions. Each option, according to the classical view, is assigned a subjective value, or "utility," and the person then proceeds to choose the option assigned the highest utility. A direct consequence of this account is that offering more alternatives is always a good thing, since the more options there are, the more likely is the consumer to find one that proves sufficiently attractive.

In contrast to this model, behavioral research suggests that, since preferences tend to be constructed in the context of decision, choices can prove difficult to make. People often search for a compelling rationale for choosing one option over another. Whereas sometimes a compelling reason can be articulated, at other times no easy rationale presents itself, rendering the conflict between options hard to resolve. Such conflict can lead to the postponing of decision or to a passive resort to a "default" option and can generate preference patterns that are fundamentally different from those predicted by accounts based on value maximization. In particular, the addition of options can excessively complicate (and, thus, "worsen") the offered set, whereas the normative rational choice assumption is that added options only make things better (Iyengar and Lepper, 2000; Shafir, Simonson, and Tversky, 1993; Tversky and Shafir, 1992).

In one study, for example, expert physicians had to decide about medication for a patient with osteoarthritis. These physicians were more likely to decline prescribing a new medication when they had to choose between two new medications than when only one new medication was available (Redelmeier and Shafir, 1995). The difficulty of choosing between the two medications presumably led some physicians to recommend not starting either. A similar pattern was documented with shoppers in an upscale grocery store, where tasting booths offered the opportunity to taste 6 different jams in one condition, or any of 24 jams in the second condition. Of those who stopped to taste, 30% proceeded to purchase a jam in the 6-jams condition, whereas more stopped but only 3% purchased a jam in the 24-jam condition (Iyengar and Lepper, 2000). Of even greater relevance to the topic at hand, Iyengar, Jiang, and Huberman (2004) showed that employees' participation in 401(k) plans drops as the number of fund options made available by their employer increases.

Bertrand et al. (2010) conducted a field experiment with a local lender in South Africa to assess the relative importance of various subtle psychological manipulations in the decision to take up a loan offer. Clients were sent letters offering large, short-term loans at randomly assigned interest rates. In addition, several psychological features on the offer letter were also independently randomized, one of which was the number of sample loans shown: the offer letters displayed either one example of a loan size and term, along with respective monthly repayments, or it displayed four such examples. In contrast with standard economic thinking and in line with conflict-based predictions, higher take-up was observed under the one-option description than under the multiple-options version. The magnitude of this effect was

large: relative to the multiple-options version, the single-option description had the same positive effect on take-up as dropping the monthly interest on these loans by more than 2 percentage points.

Mental Accounting

In their intuitive mental accounting schemes, people compartmentalize wealth and spending into distinct budget categories, such as savings, rent, and entertainment, and into separate mental accounts, such as current income, assets, and future income (Thaler, 1985; 1992). Contrary to standard fungibility assumptions, people exhibit different degrees of willingness to spend from their diverse accounts. Compartmentalization can serve useful functions in managing one's behavior, but it also can yield consumption patterns that are overly dependent on current income and sensitive to labels, which can lead to saving (at low interest rates) and borrowing (at higher rates) at the same time (Ausubel, 1991).

An understanding of such proclivities may help firms design instruments that bring about more desirable outcomes. For instance, with respect to retirement saving, the tendency to spend one's savings is lower when monies are not in transaction accounts. And faulty planning, distraction, and procrastination all account for the persistent findings that saving works best as a default. Participation in 401(k) plans is significantly higher when employers offer automatic enrollment (Madrian and Shea, 2001), and because participants tend to retain the default contribution rates and have an easier time committing now to a costly step in the future, savings can be increased as a result of agreeing to increased deductions from future raises (Benartzi and Thaler, 2004; see also Benartzi, Peleg, and Thaler, this volume).

Knowledge and Attention

Standard theory assumes that consumers are attentive and knowledgeable and typically able to gauge and avail themselves of important information. In contrast, research suggests that many individuals lack knowledge of relevant options, program rules, benefits, and opportunities, and not only among the poor or the uneducated. Surveys show that less than one-fifth of investors (in stocks, bonds, funds, or other securities) can be considered "financially literate" (Alexander, Jones, and Nigro, 1998), and similar findings describe the understanding shown by pension plan participants (Schultz, 1995). Indeed, even older beneficiaries often do not know what kind of pension they are set to receive, or what mix of stocks and bonds are held in their retirement accounts (Lusardi, Mitchell, and Curto, 2009).

The amount of information people can and do attend to is limited. Moreover, cognitive load has been shown to affect performance in everyday tasks. To the extent that consumers find themselves in challenging situations that are unfamiliar, tense, or distracting, all of which consume cognitive resources, less focused attention will be available to process the information that is relevant to the decision at hand. This, in turn, can render decision making even more dependent on situational cues and peripheral considerations, all the more so for "low literate" participants, who tend to experience even greater difficulties with effort versus accuracy trade-offs, show overdependence on peripheral cues, and tend toward a systematic withdrawal from many market interactions (Adkins and Ozanne, 2005).

Information cannot be thought of as naturally yielding knowledge, and knowledge cannot be assumed to generate the requisite behavior. People often do not fully process data that is imminently available because of limitations in attention, understanding, perceived relevance, misremembering, or misforecasting its impact. This is often underappreciated by program designers, who tend to believe that people will know that which is important and knowable. In summary, for participants with limited cognitive resources—whose decisions are heavily dependent on insufficient knowledge, perceived norms, automatic defaults, and other minor contextual nuances—regulation merits even greater attention with regard to nuanced behavioral factors.

The Power of Institutions

The substantial influence of context on behavior implies, among other things, that institutions will come to play a central role in shaping how people think and what they do. By institutions, we mean formal laws and rules, firms and other organizations, structures and governments, and widespread market practices (see, e.g., Sherraden and Barr, 2005). Among other things,

1. Institutions shape defaults, the "favored" starting point. It is now well established that defaults can have a profound influence on the outcomes of individual choices. Data available on decisions ranging from retirement savings and portfolio choices to the decision to be a willing organ donor illustrate the substantial increase in market share of default options (Johnson and Goldstein, 2003; Johnson et al., 1993; see in this volume, Johnson and Goldstein; Benartzi, Peleg, and Thaler). Contrary to a view where the

default is just one of a number of alternatives, in reality defaults persist. This persistence not only stems from confusion about available options, procrastination, forgetting, and other sources of inaction, but also may be fostered because the default is perceived as the most popular option (often a self-fulfilling prophecy), is implicitly recommended by experts, or is endorsed by the government.

2. Institutions shape behavior. Many low-income families are, de facto, savers, whether or not they resort to banks. But the availability of institutions to help foster savings can make a big difference (Barr, 2004; Berry, 2004). Without the help of a financial institution, people's savings are at risk (including from theft, impulse spending, and the needs of other household members), savings grow more slowly, and they may not be available as an emergency cushion or to support access to reasonably priced credit in times of need. Institutions provide safety, guidance, and control. In circumstances of momentary need, temptation, distraction, or limited self-control, those savers who are unbanked are likely to find it all the more difficult to succeed on the path to long-term financial stability.

Consider, for example, two individuals with no access to credit cards: one has her paycheck directly deposited into a savings account, and the other does not. Whereas cash is not readily available to the first person, who needs to take active steps to withdraw it, cash is immediately available to the second, who must take active measures to save it. The greater tendency to spend cash in the wallet compared to funds deposited in the bank (Thaler, 1999) suggests that the first, banked person will spend less on impulse and save more easily than the person who is unbanked. Holding risk- and savings-related propensities constant, the first person is likely to end up a more active and efficient saver than the second, due to nothing but a seemingly minor institutional arrangement.

Direct deposit is an institution that can have a profound effect on saving. A recent survey conducted by the American Payroll Association (2002) suggests that American employees are gaining confidence in direct deposit as a reliable method of payment that gives them greater control over their finances, and that employers are recognizing direct deposit as a low-cost employee benefit that can also save payroll processing time and money. The employers of the poor, in contrast, often do not require nor propose electronic salary payments. Instead, they prefer not to offer direct deposit to hourly/nonexempt employees, temporary or seasonal employees, part-timers, union employees, and employees in remote locations, all categories that correlate with being low paid. The most frequently stated reasons for not offering direct deposit to these employees include lack of processing time to meet standard industry (Automated Clearing House) requirements, high turnover, and union contract restrictions. All this constitutes a missed opportunity to offer favorable access to direct deposit for needy individuals, whose de facto default consists of going after hours to cash their modest check for a hefty fee.

3. Institutions provide implicit planning. As it turns out, a variety of institutions provide implicit planning, often in ways that address potential behavioral weaknesses. Credit card companies send customers timely reminders of due payments, and clients can elect to have their utility bills automatically charged, allowing them to avoid late fees if occasionally they do not get around to paying in time. The low-income buyer, on the other hand, without the credit card, the automatic billing, or the web-based reminders, risks missed payments, late fees, disconnected utilities (followed by high reconnection charges), etc. In fact, institutions can also sabotage planning, for example, by providing debt too easily. Temporal discounting in general and present bias in particular can be exploited to make immediate cash more attractive than any menacing future costs.

A behavioral analysis yields new appreciation for the impact of institutions, which affect people's lives by, among other things, easing their planning, helping them transform their intentions into actions, or enabling their resistance to temptation. Consider again the case of a low-income household. Having little slack, low-income households cannot readily cut back consumption in the face of an unanticipated need or shock (Mullainathan and Shafir, 2009). When they do cut back, it is often on essentials. In many instances, cutting back means paying late, and paying late means incurring costly late fees, utility or phone reconnection fees (Edin and Lein, 1997), and serious disruptions to work, education, and family life. In other cases it means costly short-term borrowing to avoid those consequences. In principle, the lack of slack should provide low-income households a strong incentive to increase their buffer-stock savings to cope with a volatile environment. Yet such households tend to have negligible liquid savings, in part because the financial system makes it difficult for them to get access to affordable savings vehicles (Barr, 2004).

Financial services may provide an important pathway out of poverty. Such services facilitate savings to mitigate shocks and promote asset development, and they facilitate borrowing to purchase higher-cost durables or to help weather tough times. In short, financial services allow individuals to smooth consumption and invest. Improvement of financial services, then,

provides two key advantages. First, for individuals who have access to financial services, improvement would lower the costs they pay. For example, improved financial services may enable them to use a credit card rather than a more expensive payday lender. Second, individuals who have not had access to financial services would get the direct benefits of access, such as the ability to borrow to smooth shocks (e.g., an illness, job loss, or divorce).

Access to financial institutions allows people to improve their planning by keeping money out of temptation's way. Direct deposit and automatic deductions can remove the immediate availability of cash and put in place automatic savings. Financial institutions can make it easy for individuals to make infrequent, carefully considered financial accounting decisions that can prove resistant to occasional intuitive error or to momentary impulse. In this sense, improved financial institutions can have a disproportionate impact on the lives of the poor. Moving from a payday lender and a check casher to a bank with direct deposit and payroll deduction can have benefits in the form of improved planning, saving, temptation avoidance, and other outcomes far more important than the transaction costs saved.

Behavior, Markets, and Policy: A Conceptual Framework

A behavioral perspective provides a better account of how individuals make decisions and is thus a useful corrective to the rational agent model. Yet a model focused on individuals is, on its own, incomplete as a basis for policy. The perspective outlined above needs to be embedded in the logic of markets. A framework is required that takes into account firms' incentives with respect to individual behavior as well as to regulation. This perspective produces two dimensions to consider: firms' interactions with consumers, and firms' interactions with regulators.

As for the first, the psychological biases of individuals can either be aligned with, or in opposition to, the interest of firms that market products or services. Consider a consumer who does not fully appreciate the profound effects of the compounding of interest. This consumer would be prone both to undersave and to overborrow. And both the consumer and society would prefer that he did not have such a bias in both contexts. Firms, for their part, would also prefer that the individual not have the bias to undersave, so that funds intended for investment and for fee generation would not diminish (abstracting from fee structures). However, at least over the short term, firms would be perfectly content to see the same individual overborrow (abstracting from collection costs).

With regard to the second dimension, the market response to individual failure can profoundly affect regulation. In attempting to boost participation in 401(k) retirement plans, for example, the regulator faces at worst indifferent and at best positively inclined employers and financial firms.[1] With respect to credit, by contrast, firms often have strong incentives to *exacerbate* psychological biases by failing to highlight the costs of borrowing. Regulation in this case faces a much more difficult challenge than in the savings situation. In forcing the disclosure of hidden prices of credit, the regulator often faces uncooperative firms, whose interests are to find ways to work around or undo regulatory interventions.

The mode of regulation chosen should take account of this interaction between firms and individuals and between firms and regulators. One might think of the regulator as holding two kinds of levers, which we describe as changing the rules and changing the scoring.[2] When forcing disclosure of the APR, for example, the regulator effectively changes the "rules" of the game: what a firm must say. A stronger form of rule change is product regulation: changing what a firm must do. Behavioral rule changes, such as creating a favored starting position or default, fall somewhere in between. When imposing liability, by contrast, the regulator changes the way the game is "scored." Liability levels can be set, in theory, to match or exceed the gains to the firm from engaging in the disfavored activity. Scoring can also be changed, for example, by providing tax incentives to engage in the favored activity or by imposing negative tax consequences for engaging in a disfavored activity. Typically, changing the rules of the game (without changing the scoring) alters certain behaviors while maintaining the firms' original incentives; changing the scoring of the game can alter those incentives.

Understanding the interaction between individuals, firms, and regulators in particular markets highlights the care that must be taken when transferring behavioral economic insights from one domain to another. For example, the insights of the most prominent example of behavioral regulation—setting defaults in 401(k) participation—ought not to be mindlessly applied to other markets. Changing the rules on retirement saving by introducing defaults works well because employers' incentives align (or do not misalign) with regulatory efforts to guide individual choice. In other words, under current conditions, employers are unaffected, or may even be hurt, by individuals' propensity to undersave in 401(k) plans.[3] Consequently, they will not lean against attempts to fix that problem.

In other instances, where firms' incentives misalign with regulatory intent, changing the rules alone may not work since those firms have strong incentives to work creatively around those changes. Interestingly, such circumstances may lead to regulations, such as "changing the scoring" with liability, which, although deeply motivated by behavioral considerations, are not themselves particularly psychological in nature. That is, given market responses, rules based on subtle attempts to influence individual psychology, for example through defaults or framing, may be too weak, and changes in liability rules or other measures may prove necessary.

The distinction in market responses to individual psychology is central to our framework and is illustrated in table 26.1. In some cases, the market is either neutral or wants to overcome consumer fallibility. In other cases, the market would like to exploit or exaggerate consumer fallibility. When consumers misunderstand compounding of interest in the context of *saving*, banks have incentives to reduce this misunderstanding so as to increase deposits. When consumers misunderstand compounding in the context of *borrowing*, lenders may lack the incentive to correct this misunderstanding because they can induce consumers to overborrow in ways that maintain or enhance profitability, at least over market-relevant time horizons.[4] When consumers procrastinate in signing up for the Earned Income Tax Credit (and hence in filing for taxes), private tax preparation firms have incentives to help remove this procrastination so as to increase their customer base. When consumers procrastinate in returning rebates (but make retail purchases intending to get a rebate), retailers benefit. Note the parallelism in these examples: firms' incentives to alleviate or exploit a bias are not an intrinsic feature of the bias itself. Instead, they are a function of how the bias plays itself out in the particular market structure.

In the consumer credit market, one worries that many interactions between individuals and firms are of the kind where firms seek to exploit, rather than alleviate, bias. If true, this raises the concern of overextrapolating from the 401(k) defaults example to credit products. To the extent that 401(k) defaults work because the optimal behavior is largely aligned with market incentives, other areas, such as credit markets, might be more difficult to regulate with mere defaults. Furthermore, if the credit market is dominated by "low-road" firms offering opaque products that "prey" on human weakness, it is more likely that regulators of such a market will be captured because "high road" interests with small market share will tend to be too weak politically to push back against the bigger low-road players. Market forces will then defeat weak positive interventions, such as the setting of defaults, and low-road players will continue to dominate. Many observers, for example, believe that credit card markets were, at least prior to passage of the CARD Act in 2009, dominated by such low-road practices (see, e.g., Bar-Gill, 2004; Mann, 2007). If government policy makers want to attempt to use defaults in such contexts, they might need to deploy "stickier" defaults (namely, ones that might prove costly to abandon) or other more aggressive policy options.

In our conceptual approach to the issue of regulatory choice (table 26.2), the regulator can either change the rules of the game or change the scoring of the game. Setting a default is an example of changing the rules of the game, as is disclosure regulation. The rules of the game are changed when there is an attempt to change the nature of the interactions between individuals and firms, as when the regulation attempts to affect what can be said, offered, or done. Changing the scoring of the game, by contrast, changes the payoffs a firm will receive for particular outcomes. This may be done without a particular rule

Table 26.1 The firm and the individual

Behavioral fallibility	Market neutral and/or wants to overcome consumer fallibility	Market exploits consumer fallibility
Consumers misunderstand compounding	Consumers misunderstand compounding in *savings*	Consumers misunderstand compounding in *borrowing*
	→ Banks would like to *reduce* this to increase savings base	→ Banks would like to *exploit* this to increase borrowing
Consumers procrastinate	Consumers procrastinate in signing up for EITC	Consumers procrastinate in returning rebates
	→ Tax filing companies would like to *reduce* this so as to increase number of customers	→ Retailers would like to *exploit* this to increase revenues

Table 26.2 Changing the game

Rules	Set the defaults in 401(k) savings
	Opt-out rule for organ donation
Scoring	Penalties for 401(k) enrollment top heavy with high-salary employees
	Grants to states that enroll organ donors

about how the outcome is to be achieved. For example, pension regulation that penalizes firms whose 401(k) plan enrollment is top-heavy with high-paid executives is an example of how scoring gives firms *incentives* to enroll low-income individuals without setting particular rules on how this is done. Changing rules and changing scoring often accompany each other, but they are conceptually distinct.

Table 26.3 weaves these approaches together, illustrating our conceptual framework for behaviorally informed regulation. The table shows how regulatory choice may be analyzed according to the market's stance toward human fallibility. On the left side of the table, market incentives align reasonably well with society's goal of overcoming consumer fallibility. Rules in that context may have a relatively lighter touch. For example, using automatic savings plans as a default in retirement saving, or providing for licensing and registration to ensure that standard practices are followed. Similarly, scoring on the left side of the table might involve tax incentives to reduce the costs to firms of engaging in behaviors that align well with their market interests and the public interest but may otherwise be too costly. On the right side of the table, by contrast, market incentives are largely misaligned with the public interest in overcoming consumer fallibility. In that context, rule changes will typically need to be more substantial to be effective and may need to be combined with changing the scoring.

Table 26.3 Behaviorally informed regulation

	Market neutral and/or wants to overcome consumer fallibility	Market exploits consumer fallibility
Rules	Public education on saving	Sticky defaults (opt-out mortgage or credit card)
	Direct deposit/ auto-save	Information debiasing (payoff time and cost for credit cards)
	Licensing	
Scoring	Tax incentives for savings vehicles Direct deposit tax-refund accounts	*Ex post* liability standard for truth in lending Broker duty of care and changing compensation practices (yield spread premiums)

The discussion that follows illustrates the challenge to policies in the top right-hand corner of table 26.3. Changing the rules of the game alone will often be insufficient when firms are highly motivated to find work-arounds. As such, merely setting a default—in contrast to defaults deployed in markets on the left side of the table—will likely not work. Thus, when we suggest opt-out policies in mortgages below, the challenge will be to find ways to make these starting positions "sticky" so that firms do not easily undo their default nature. In such cases, achieving an effective default may require separating low-road from high-road firms and making it profitable for high-road firms to offer the default product (for a related concept, see Kennedy, 2005). For that to work, the default must be sufficiently attractive to consumers, sufficiently profitable for high-road firms to succeed in offering it, and the penalties associated with deviations from the default must be sufficiently costly so as to make the default stick even in the face of market pressures from low-road firms. In some credit markets, low-road firms may become so dominant that sticky defaults will be ineffectual. Moreover, achieving such a default is likely to be costlier than making defaults work when market incentives align, not least because the costs associated with the stickiness of the default involve greater dead-weight losses due to higher costs to opt out for those for whom deviating from the default is optimal. Such losses would need to be weighed against the losses from the current system, as well as against losses from alternative approaches, such as disclosure or product regulation. Nonetheless, given the considerations above, it seems worth exploring whether sticky defaults can help to transform consumer financial markets in certain contexts.

Sticky defaults are one of a set of examples we discuss as potential regulatory interventions based on our proposed conceptual framework. As noted above, given market responses to relevant psychological factors in different contexts, regulation may need to take a variety of forms, including some that, while perhaps informed by psychology, are designed not to affect behavioral change but rather to alter the market structure in which the relevant choices are made. Given the complexities involved, our purpose here is not to champion specific proposals but rather to illustrate how a behaviorally informed regulatory analysis may generate a deeper understanding of the costs and benefits of particular policies.

Behaviorally Informed Financial Regulation

Following Barr, Mullainathan, and Shafir (2008a), we review a set of ideas to illustrate our conceptual framework in three main areas of consumer finance: home mortgages, credit cards, and bank accounts. We will use these three substantive areas to explore how changing the rules and changing the scoring can affect firms' behavior in market contexts where firms have incentives to exploit consumer bias (as in credit) and in those where firms have incentives to overcome such biases (as in saving). Our analyses map into different quadrants of table 26.3. Since we first published our work, there has been significant progress in implementing a number of these ideas.[5] We therefore also discuss how some of these ideas have been recently implemented in the CARD Act of 2009, the Dodd-Frank Wall Street and Consumer Protection Act of 2010, and other policy initiatives. In addition, with the creation of the new Consumer Financial Protection Bureau (CFPB) in the Dodd-Frank Act, there is an opportunity to further learn from behavioral research and to experiment with new approaches. We will briefly highlight some of these opportunities.

Behaviorally Informed Home Mortgage Regulation

FULL INFORMATION DISCLOSURE TO DEBIAS BORROWERS

With the advent of nationwide credit reporting systems and the refinement of credit scoring and modeling, the creditor and broker know information about the borrower that the borrower does not necessarily know about himself, including not just his credit score, but his likely performance regarding a particular set of loan products. Creditors will know whether the borrower could qualify for a better, cheaper loan, as well as the likelihood that he will meet his obligations under the existing mortgage or become delinquent, refinance, default, or go into foreclosure. Yet lenders are not required to reveal this information to borrowers, and the impact of this lack of disclosure is probably exacerbated by consumer beliefs. Consumers likely have false background assumptions regarding what brokers and creditors reveal and the implications of their offers. What if consumers believe the following:

> Creditors reveal all information about me and the loan products I am qualified to receive. Brokers work for me in finding me the best loan for my purposes, and lenders offer me the best loans for which I qualify. I must be qualified for the loan I have been offered, or the lender would not have validated the choice by offering me the loan. Because I am qualified for the loan that must mean that the lender thinks that I can repay the loan. Why else would they lend me the money? Moreover, the government tightly regulates home mortgages; they make the lender give me all these legal forms. Surely the government must regulate all aspects of this transaction.

In reality, the government does not regulate as the borrower believes, and the lender does not necessarily behave as the borrower hopes. Instead, information is hidden from the borrower, information that would improve market competition and outcomes. Given the consumer's probably false background assumptions and the reality of asymmetric information favoring the lender and broker, we suggest that creditors be required to reveal useful information to the borrower at the time of the mortgage loan offer, including disclosure of the borrower's credit score and the borrower's qualifications for the lender's mortgage products and rates. Such an approach corresponds to the provision of debiasing information, in the top right of table 26.3.

The goal of these disclosures would be to put pressure on creditors and brokers to be honest in their dealings with applicants. The additional information might improve comparison shopping and, perhaps, outcomes. Of course, revealing such information would also reduce brokers' and creditors' profit margins. But because the classic market competition model relies on full information and assumes rational behavior based on understanding, this proposal simply attempts to remove market frictions from information failures and to move market competition more toward its ideal. By reducing information asymmetry, full information disclosure would help debias consumers and lead to more competitive outcomes.

EX POST STANDARDS-BASED TRUTH IN LENDING

Optimal disclosure will not occur in all markets through competition alone because in many contexts firms have incentives to hide information about products or prices and because consumers will not insist on competition based on transparency due to a lack of knowledge or understanding and a misforecasting of their own behavior. Competition under a range of plausible scenarios will not necessarily generate psychologically informative and actionable disclosure. Moreover, even if all firms have an incentive to disclose in meaningful ways, they may not disclose in the same way, thus undermining comparison shopping by consumers. If competition does not produce informative disclosure, disclosure regulation might be necessary. But the mere fact that disclosure regulation is needed does not mean that it will work.

A behavioral perspective should focus in part on improving the disclosures themselves. The goal of disclosure should be to improve the quality of information about contract terms in meaningful ways. Simply adding information, for example, is unlikely to work. Disclosure policies are effective to the extent that they present a frame—a way of parsing the disclosure—that is both well understood and conveys salient information that helps the decision maker act optimally. It is possible, for example, that information about the failure frequency of particular products ("2 out of 10 borrowers who take this kind of loan default") might help, but proper framing can be difficult to achieve and to maintain consistently, given that it may vary across situations. Moreover, the attempt to improve decision quality through better consumer understanding, which is presumed to change consumers' intentions, and consequently their actions, is fraught with difficulty. There is often a wide divide between understanding, intention, and action.

Furthermore, even if meaningful disclosure rules can be created, sellers can generally undermine whatever ex ante disclosure rule is established, in some contexts simply by "complying" with it: "Here's the disclosure form I'm supposed to give you, just sign here." With rules-based, ex ante disclosure requirements, the rule is set up first, and the firm (the discloser) moves last. While an ex ante rule may attempt to provide information and facilitate comparison shopping, whatever incentives the discloser had to confuse consumers persist in the face of the regulation. While officially complying with the rule, there is market pressure to find other means to avoid the salutary effects on consumer decisions that the disclosure was intended to achieve.

In light of the challenges inherent to addressing such issues ex ante, we propose that policy makers consider shifting away from sole reliance on a rules-based, ex ante regulatory structure for disclosure as embodied in the Truth in Lending Act (TILA) and toward the integration of an ex post, standards-based disclosure requirement in addition. Rather than sole reliance on a rule, we would also deploy a standard, and rather than sole reliance on an ex ante decision about content, we would permit the standard to be enforced after the event, for example, after loans are made. In essence, courts or the new CFPB would determine whether the disclosure would have, under common understanding, effectively communicated the key terms of the mortgage, conforming to some minimum standard, to the typical borrower. This approach could be similar to ex post determinations of reasonableness of disclaimers of warranties in sales contracts under UCC 2-316 (Uniform Commercial Code; see White and Summers, 1995). This type of policy intervention would correspond to a change in "scoring," as in the lower right of table 26.3.

An ex post version of truth in lending based on a reasonable-person standard to complement the fixed disclosure rule under TILA might permit innovation—both in products themselves and in disclosure—while minimizing rule evasion. An ex post standard with sufficient teeth could change the incentives of firms to confuse and would be more difficult to evade. Under the current approach, creditors can easily "evade" TILA by simultaneously complying with its actual terms while making the required disclosures of the terms effectively useless in the context of borrowing decisions by consumers with limited attention and understanding. TILA, for example, does not block a creditor from introducing a more salient term ("lower monthly cost!") to compete with the disclosed APR for borrowers' attention. By contrast, under an ex post standards approach, lenders could not plead mere compliance with a TILA rule as a defense. Rather, the question would be one of objective reasonableness: whether the lender meaningfully conveyed the information required for a typical consumer to make a reasonable judgment about the loan. Standards would also lower the cost of specification ex ante. Clarity of contract is hard to specify *ex ante* but easier to verify *ex post*. Over time, through agency action, guidance, model disclosures, no-action letters, and court decisions, the parameters of the reasonableness standard would become known and predictable.

While TILA has significant shortcomings, we do not propose abandoning it. Rather, TILA would remain and could be improved with a better understanding of consumer behavior. The Federal Reserve Board, for example, unveiled major and useful changes to its disclosure rules based on consumer research.[6] TILA would still be important in setting uniform rules to permit comparison shopping among mortgage products, one of its two central goals. However, some of the burden of TILA's second goal, to induce firms to reveal information that would promote better consumer understanding even under circumstances in which the firm believes that it would hurt the firm, would be shifted to the ex post standard.

There would be significant costs to such an approach, especially at first. Litigation or regulatory enforcement would impose direct costs, and the uncertainty surrounding enforcement of the standard ex post might deter innovation in the development of mortgage products. The additional costs of compliance with a disclosure standard might reduce lenders' willingness to develop new mortgage products designed to reach lower-income or minority borrowers who might not be served by the firms' "plain vanilla" products.[7] The lack of clear rules might also increase

consumer confusion regarding how to compare innovative mortgage products to each other, even while it increases consumer understanding of the products being offered. Ultimately, if consumer confusion results mostly from firm obfuscation, then our proposal will likely help a good deal. By contrast, if consumer confusion in this context results mostly from market complexity in product innovation, then the proposal is unlikely to make a major difference and other approaches focused on loan comparisons might be warranted (see, e.g., Thaler and Sunstein, 2008, this volume).

Despite the shortcomings of an ex post standard for truth in lending, we believe that such an approach is worth pursuing. To limit the costs associated with our approach, the ex post determination of reasonableness could be significantly confined. For example, if courts are to be involved in enforcement, the ex post standard for reasonableness of disclosure might be limited to providing a (partial) defense to full payment in foreclosure or bankruptcy, rather than being open to broader enforcement through affirmative suits for damages. Alternatively, rather than court enforcement, the ex post standard might be applied solely by the CFPB through supervision. Furthermore, the ex post exposure might be significantly reduced through ex ante steps. For example, the CFPB might develop safe harbors for reasonable disclosures, issue model disclosures, or use no-action letters to provide certainty to lenders. Moreover, firms might be tasked with conducting regular surveys of borrowers or conducting experimental design research to validate their disclosures; results from the research demonstrating a certain level of consumer understanding might provide a rebuttable presumption of reasonableness or even a safe harbor from challenge.[8] The key is to give the standard sufficient teeth without deterring innovation. The precise contours of enforcement and liability are not essential to the concept, and weighing the costs and benefits of such penalties, as well as detailed implementation design, are beyond the scope of introducing the idea here.

STICKY OPT-OUT MORTGAGE REGULATION

While the causes of the mortgage crisis are myriad, a central problem was that many borrowers took out loans that they did not understand and could not afford. Brokers and lenders offered loans that looked much less expensive than they really were, because of low initial monthly payments and hidden, costly features. Families commonly make mistakes in taking out home mortgages because they are misled by broker sales tactics, misunderstand the complicated terms and financial tradeoffs in mortgages, wrongly forecast their own behavior, and misperceive their risk of borrowing. How many homeowners really understand how the teaser rate, introductory rate, and reset rate relate to the London Interbank Offered Rate plus some specified margin, or how many can judge whether the prepayment penalty will offset the gains from a teaser rate?

Altering the rules of the game of disclosure, and altering the "scoring" for seeking to evade proper disclosure, may be sufficient to reduce the worst outcomes. However, if market pressures and consumer confusion are sufficiently strong, such disclosure may not be enough. If market complexity is sufficiently disruptive to consumer choice, product regulation might prove most appropriate. For example, by barring prepayment penalties, one could reduce lock-ins to bad mortgages; by barring short-term ARMs and balloon payments, one could reduce the pressure to refinance; in both cases, more of the cost of the loan would be pushed into interest rates, and competition could focus on an explicitly stated price in the form of the APR. Such price competition would benefit consumers, who would be more likely to understand the terms on which lenders were competing. Product regulation would also reduce cognitive and emotional pressures related to potentially bad decision making by reducing the number of choices and eliminating loan features that put pressure on borrowers to refinance on bad terms. However, product regulation may stifle beneficial innovation, and there is always the possibility that the government may simply get it wrong, prohibiting good products and permitting bad ones.

For that reason, we proposed a new form of regulation.[9] We proposed that a default be established with increased liability exposure for deviations that harm consumers. For lack of a better term, we called this a sticky opt-out mortgage system. A sticky opt-out system would fall, in terms of stringency, between product regulation and disclosure. For reasons we will explain below, market forces would likely swamp a pure opt-out regime—that is where the need for stickiness came in. This approach corresponds to a combination of changing the rules of the game (top right of table 26.3), and changing liability standards (bottom right of that table).

The proposal is grounded in our equilibrium model of incentives for firms and of individual psychology. Many borrowers may be unable to compare complex loan products and act optimally for themselves based on such an understanding (see, e.g., Ausubel, 1991). We thus deploy an opt-out strategy to make it easier for borrowers to choose a standard product and harder for them to choose a product they are less likely to understand. At the same time,

lenders may seek to extract surplus from borrowers because of asymmetric information about future income or default probabilities (Musto, 2007), and, in the short term, lenders and brokers may benefit from selling borrowers loans they cannot afford. Thus, a pure default would be undermined by the firms, and regulation needs to take account of this market pressure by pushing back.

In our model, lenders would be required to offer eligible borrowers a standard mortgage (or set of mortgages), such as a fixed-rate, self-amortizing thirty-year mortgage loan or a standard ARM product according to reasonable underwriting standards. The precise contours of the standard set of mortgages would be set by regulation. Lenders would be free to charge whatever interest rate they wanted on the loan and, subject to the constraints outlined below, could offer whatever other loan products they wanted outside of the standard package. Borrowers, however, would get the standard mortgage offered, unless they chose to opt out in favor of a nonstandard option offered by the lender, after honest and comprehensible disclosures from brokers or lenders about the terms and risks of the alternative mortgages. An opt-out mortgage system would mean borrowers would be more likely to get straightforward loans they could understand.

But a plain-vanilla opt-out policy is likely to be inadequate. Unlike the savings context, where market incentives align well with policies to overcome behavioral biases, in the context of credit markets, firms often have an incentive to hide the true costs of borrowing. Given the strong market pressures to deviate from the default offer, we would need to require more than a simple opt-out to make the default stick. Deviation from the offer would require heightened disclosures and additional legal exposure for lenders in order to make the default sticky. Under our plan, lenders would have stronger incentives to provide meaningful disclosures to those whom they convince to opt out, because they would face increased regulatory scrutiny or increased costs if the loans did not work out.

Future work will need to explore in greater detail the enforcement mechanism. For example, under one potential approach to making the opt-out sticky, if default occurs after a borrower has opted out, the borrower could raise the lack of reasonable disclosure as a defense to bankruptcy or foreclosure. Using an objective reasonableness standard akin to that used for warranty analysis under the Uniform Commercial Code,[10] if the court determined that the disclosure would not effectively communicate the key terms and risks of the mortgage to the typical borrower, the court could modify the loan contract. Although Congress rejected this proposal in the Dodd-Frank Act, if Congress were to revisit the issue, it could authorize the CFPB to enforce the requirement on a supervisory basis rather than relying on the courts. The agency would be responsible for supervising the disclosures according to a reasonableness standard and would impose a fine on the lender and order corrective actions if the disclosures were found to be unreasonable. The precise nature of the stickiness required and the trade-offs involved in imposing these costs on lenders would need to be explored in greater detail, but in principle, a sticky opt-out policy could effectively leverage the behavioral insight that defaults matter with the industrial organizational insight that market incentives work against the advantages of a pure opt-out policy in many credit markets.

An opt-out mortgage system with stickiness might provide several benefits over current market outcomes. For one, a "plain vanilla" set would be easier to compare across mortgage offers. Information would be more efficiently transmitted across the market. Consumers would be likely to understand the key terms and features of such standardized products better than they would alternative mortgage products. Price competition would be more salient once the features were standardized. Behaviorally, when alternative, "non-vanilla" products are introduced, the consumer would be made aware that these represent deviations from the default, anchoring consumers on the default product and providing some basic expectations for what ought to enter into the choice. Framing the mortgage choice as one between accepting standard mortgage offers and needing affirmatively to choose a nonstandard product should improve consumer decision making. Creditors will be required to make heightened disclosures about the risks of alternative loan products, subject to legal sanction in the event of failure to reasonably disclose such risks; the legal sanctions should deter creditors from making highly unreasonable alternative offers with hidden and complicated terms. Consumers may be less likely to make significant mistakes. In contrast to a pure product regulation approach, the sticky default approach allows lenders to continue to develop new kinds of mortgages, but only when they can adequately explain key terms and risks to borrowers.

Moreover, requiring a default accompanied by heightened disclosures and increased legal exposure for deviations may help boost high-road lending relative to low-road lending—at least if deviations resulting in harm are appropriately penalized. If offering an opt-out mortgage product helps to split the market between high- and low-road firms and rewards the former, the market may shift (back) toward firms that offer home mortgage products that better serve borrowers. For this to work effectively, the default—and efforts to make it sticky—should enable the consumer

easily to distinguish the typical "good" loan, benefiting both lender and borrower, from a wide range of "bad" loans that benefit the lender with higher rates and fees but harm the borrower; that benefit the borrower but harm the lender; or that harm borrower and lender but benefit third parties, such as brokers.

There will be costs associated with requiring an opt-out home mortgage. For example, sticky defaults may not be sticky enough to alter outcomes, given market pressures. The default could be undermined through the firm's incentive structures for loan officers and brokers, which could provide greater rewards for nonstandard loans. Implementation of the measure may be costly, and the disclosure requirement and uncertainty regarding enforcement of the standard might reduce overall access to home mortgage lending. There may be too many cases in which alternative products are optimal, so that the default product is in essence "incorrect" and comes to be seen as such. The default would then matter less over time, and the process of deviating from it would become increasingly just another burden (like existing disclosure paperwork) along the road to getting a home mortgage loan. Low-income, minority, or first-time homeowners who have benefited from more flexible underwriting and more innovative mortgage developments might see their access reduced if the standard set of mortgages does not include products suitable to their needs.

One could improve these outcomes in a variety of ways. For example, the opt-out regulation could require that the standard set of mortgages include a thirty-year fixed mortgage, a five- or seven-year adjustable-rate mortgage, and straightforward mortgages designed to meet the particular needs of first-time, minority, or low-income homeowners. One might develop "smart defaults," based on key borrower characteristics, such as income and age. With a handful of key facts, an optimal default might be offered to an individual borrower. The optimal default would consist of a mortgage or set of mortgages that most closely align with the set of mortgages that the typical borrower with that income, age, and education would prefer. For example, a borrower with rising income prospects might appropriately be offered a five-year adjustable rate mortgage. Smart defaults might reduce error costs associated with the proposal and increase the range of mortgages that can be developed to meet the needs of a broad range of borrowers, including lower-income or first-time homeowners; however, smart defaults may add to consumer confusion. Even if the consumer (with the particular characteristics encompassed by the smart default) faces a single default product, spillover from options across the market may make decision making more difficult. Finally, it may be difficult to design smart defaults consistent with fair lending rules.

If Congress were to revisit this proposal in the future, it could authorize the CFPB to implement such a program. Supervisory implementation would help to improve the standard mortgage choice set and to reduce enforcement costs over time. The CFPB could be required periodically to review the defaults and to conduct consumer experimental evaluation or survey research to test both the products and the disclosures, so that these stay current with developments in the home mortgage market. Indeed, lenders might be required to conduct such research and to disclose the results to the CFPB and the public upon developing a new product and its related disclosures. In addition, the CFPB might use the results of the research to provide safe harbors or no-action letters for disclosures that are shown to be reasonable ex ante. The CFPB could conduct ongoing supervision and testing of compliance with the opt-out regulations and disclosure requirements. Through such no-action letters, safe harbors, supervision, and other regulatory guidance, the CFPB can develop a body of law that would increase compliance across the diverse financial sectors involved in mortgage lending, while reducing the uncertainty facing lenders from the new opt-out requirement and providing greater freedom for financial innovation.

RESTRUCTURE THE RELATIONSHIP BETWEEN BROKERS AND BORROWERS

An additional approach to addressing the problem of market incentives to exploit behavioral biases would be to focus directly on restructuring brokers' duties to borrowers and reforming compensation schemes that provide incentives to brokers to mislead borrowers. Mortgage brokers have dominated the subprime market. Brokers generally have been compensated with *yield spread premiums* (YSP) for getting borrowers to pay higher rates than those for which the borrower would qualify. Such YSPs have been used widely.[11] In loans with YSPs, unlike other loans, there is a wide dispersion in prices paid to mortgage brokers. As Jackson and Burlingame (2007) have shown, within the group of otherwise comparable borrowers paying YSPs, African Americans paid $474 more for their loans, and Hispanics $590 more, than white borrowers; thus, even if minority and white borrowers could qualify for the same rate, in practice minority borrowers are likely to pay much more.[12]

Brokers cannot be monitored effectively by borrowers (See Jackson and Burlingame, 2007), and it is dubious that additional disclosures would help borrowers be better monitors (see, e.g., Federal Trade Commission, 2007), because, among other things, borrowers do not always recognize potential conflicts of interest and because brokers' disclosures of such

conflicts can paradoxically increase consumer trust (Cain, Loewenstein, and Moore, 2005). When a broker is seen to divulge that he works for himself, not in the interest of the borrower, the borrower's trust in the broker may increase: here is a broker who is being honest! Moreover, the subprime mortgage crisis suggests that while in theory creditors and investors have some incentives to monitor brokers, they do not do so effectively.

It is possible to undertake an array of structural changes regarding the broker-borrower relationship. For example, one could alter the incentives of creditors and investors to monitor mortgage brokers by changing liability rules so that broker misconduct can be attributed to lenders and creditors in suits by borrowers (see Engel and McCoy, 2007). One could directly regulate mortgage brokers through licensing and registration requirements (as is done elsewhere, e.g., in the United Kingdom); recent U.S. legislation, known as the SAFE Act, now mandates licensing and reporting requirements for brokers. In addition, the ex post disclosure standard we suggest might have a salutary effect by making it more costly for lenders when brokers evade disclosure duties, thus generating better monitoring of brokers.

We also suggest that the duties of care that mortgage brokers owe to borrowers should be raised. A higher duty of care would more closely conform to borrower expectations about the role of mortgage brokers in the market. In addition, we support the banning of YSPs that are based on the interest rate charged, for example. Banning YSPs could reduce abuses by eliminating a strong incentive for brokers to seek out higher-cost loans for customers. In fact, a number of lenders moved away from YSPs to fixed fees with some funds held back until the loan has performed well for a period of time, precisely because of broker conflicts of interest in seeking higher YSPs rather than sound loans. Banning YSPs is another way to reinforce high-road practices and protect against a renewed and profitable low-road push to increase market share once stability is restored to mortgage markets. Banning YSPs affects the payoff that brokers receive for mortgage products and thus constitutes a form of scoring change, corresponding to regulation in the bottom right of table 26.3.

PROGRESS UNDER THE DODD-FRANK ACT

The Dodd-Frank Act fundamentally reforms consumer financial protection policy in the United States. In the mortgage market, the Dodd-Frank Act undertakes a number of steps to regulate the relationship between borrowers and mortgage brokers. For example, the act requires registration and imposes a duty of care on mortgage brokers; bans steering to higher-cost products; and bans YSPs. The act requires that mortgage brokers and lenders assess a borrower's ability to repay based on documented income, taking into account the fully indexed, fully amortizing rate on a mortgage. The act prohibits mandatory predispute arbitration clauses (which limit one's right to access the courts), and it enhances disclosure requirements. It requires the use of escrow of taxes and insurance for higher-cost loans and improves escrow disclosure for all loans. It makes a number of changes to the Home Ownership and Equity Protection Act (HOEPA) to make it more effective and provide greater consumer protection.

The Dodd-Frank Act also puts in place two provisions that foster standardization in the products offered to consumers. The act requires risk retention for securitization of mortgage loans but exempts Qualified Residential Mortgages, which are designed to be standard, high-quality mortgage products with straightforward terms and solid underwriting. For loans falling outside this category that are securitized, the securitizer (or the originator) would need to retain capital to back a portion of the securitization risk. There would thus be a strong incentive to make Qualified Residential Mortgages. The Dodd-Frank Act also sets out provisions for qualified mortgages, ones for which the ability-to-pay requirement is deemed to be met. In sum, the act defines an approach to the standardization of the terms and underwriting of such mortgages. Lenders making nonqualified mortgages face a larger potential risk of liability in the event that such loans fail.

More fundamentally, the act put in place the new CFPB to supervise major financial institutions and to set rules and enforce consumer protections across the market. In addition to its authorities to set rules for and enforce existing consumer financial protection laws, the CFPB has the authority to ban unfair, deceptive, or abusive acts or practices. The bureau can also prescribe rules for disclosures of any consumer financial product. In doing so, it will rely on consumer testing, can issue model disclosures that provide a safe harbor for compliance, and may permit financial institutions to use trial disclosure programs to test out the effectiveness of alternative disclosures to those provided for in the CFPB model form. The Bureau is mandated to merge conflicting mortgage disclosures from the Real Estate Settlement Procedures Act (RESPA) and TILA into a simple form. Consumers are provided with rights to access information about their own product usage in standard, machine-readable formats. Over time, the CFPB may generate research and experimentation that will improve our understanding of consumer financial decision making, and in turn will support the bureau's supervision, rule-writing, and enforcement.

In addition to these changes to consumer financial protection, the act makes a number of changes to investor protection. For example, it provides the Securities and Exchange Commission (SEC) with authority to engage in investor testing to improve disclosures or other rules. The SEC is authorized to clarify the duties of investment advisors and broker-dealers so that they have the same high standard of care—a fiduciary duty (which, until now, investment advisors had but broker-dealers providing individualized investment advice did not). The commission is also authorized to require better disclosures of broker duties and conflicts of interest and to mandate presale disclosures for investment products. Like the CFPB, the SEC is authorized to restrict mandatory predispute arbitration. These changes should materially advance investor protections consistent with the framework we have laid out.

Behaviorally Informed Credit Card Regulation

USING FRAMING AND SALIENCE IN DISCLOSURES TO ENCOURAGE GOOD CREDIT CARD BEHAVIOR

Credit card companies have fine-tuned product offerings and disclosures in a manner that appears to be systematically designed to prey on common psychological biases—biases that limit consumer ability to make optimal choices regarding credit card borrowing (Bar-Gill, 2004). Behavioral economics suggests that consumers underestimate how much they will borrow and overestimate their ability to pay their bills in a timely manner, and credit card companies then price their credit cards and compete on the basis of these fundamental human failings. Nearly 60% of credit card holders do not pay their bills in full every month (Bucks et al., 2006). Moreover, excessive credit card debt can lead to bankruptcy (Mann, 2006). Mann (2007) has argued that credit card companies seek to keep consumers in a "sweat box" of distressed credit card debt, paying high fees for as long as possible before finally succumbing to bankruptcy.

The 2005 bankruptcy legislation focused on the need for improved borrower responsibility but paid insufficient attention to creditor responsibility for borrowing patterns.[13] Credit card companies provided complex disclosures regarding teaser rates, introductory terms, variable rate cards, penalties, and a host of other matters. Both the terms themselves and the disclosures were confusing to consumers.[14] Credit card companies, it appears, were not competing to offer the most transparent pricing.

The Office of the Comptroller of the Currency required national banks to engage in better credit card practices and to provide greater transparency on minimum payments,[15] and the Federal Reserve proposed changes to its regulations under TILA, partly in the wake of amendments contained in the bankruptcy legislation.[16] Under the proposals, for example, creditors would need to disclose that paying only the minimum balance would lengthen the payoff time and interest paid on the credit card; describe a hypothetical example of a payoff period paying only the minimum balance; and provide a toll-free number for the consumer to obtain an estimate of actual payoff time.[17] Although the very length and complexity of the board's proposal hints at the difficulty of the task of disclosure to alter consumer understanding and behavior, such improved disclosures might nevertheless help.

In earlier work (Barr, Mullainathan, and Shafir, 2008a), we proposed that Congress could require that minimum payment terms be accompanied by clear statements regarding how long it would take, and how much interest would be paid, if the customer's *actual* balance were paid off in minimum payments, and card companies could be required to state the monthly payment amount that would be required to pay the customer's actual balance in full over some reasonable period of time, as determined by regulation. These tailored disclosures use framing and salience to help consumers, whose intuitions regarding compounding and timing are weak, to make better-informed borrowing and payment choices based on their specific circumstances. Such an approach would mandate behaviorally informed changes in information disclosure rules in order to help debias consumers (corresponding to the top right of table 26.3). Although credit card companies have opposed such ideas in the past, disclosures based on the customer's actual balances are not overly burdensome, as evidenced by their implementation following the CARD Act of 2009.

Disclosures regarding the expected time to pay off actual credit card balances are designed to facilitate clearer thinking but may not be strong enough to matter. Even if such disclosure succeeds in shaping intention, we know that there is often a large gap between intention and action.[18] In fact, borrowers would need to change their behavior in the face of strong inertia and marketing by credit card companies, which often propel them to make no more than minimum payments. More generally, once enacted, market players opposed to such disclosures would promptly work to undermine them with countervailing marketing and other policies. And there could be occasional costs in other directions: for example, consumers who used to pay more than the amount required to pay off their bills in the time frame specified by regulators may now be drawn to pay off their bills more slowly. Recent preliminary research by Tufano (2009) suggests that the CARD Act may have had this

mixed effect—improving the outcomes for borrowers who paid more slowly, while worsening the outcomes for those who previously caught up more quickly than the statement's anchor on a payoff plan of three years.

AN OPT-OUT PAYMENT PLAN FOR CREDIT CARDS

A related approach, intended to facilitate behavior through intention, would be to develop an opt-out payment plan for credit cards under which consumers would need to elect a default payment level meant to pay off their existing balance over a chosen period of time unless the customer affirmatively opted out and chose an alternative payment at any point.[19] Consumers could elect to alter their chosen payment plan in advance or could, with modest friction costs, opt out and change the plan at the time they receive their bill. Such an approach corresponds to changing the rules through opt-out policies (top right of table 26.3). Given what we know about default rules, such payment plans may create expectations about consumer conduct, and in any event, inertia would cause many households simply to follow the initially chosen plan. Increasing such behavior, as driven by prior intentions, could mean lower rates of interest and fees paid, and lower incidences of financial failure. A chosen opt-out payment plan may also impose costs. Some consumers who, in the absence of the opt-out plan, would have paid off their credit cards sooner, might underestimate their capacity and opt for a slower payment plan, thus incurring higher costs from interest and fees. Alternatively, some consumers may follow their chosen opt-out payment plan when it is unaffordable for them, consequently reducing necessary consumption, such as medical care or sufficient food, or incurring other costly forms of debt. Still, confronting the need to determine a default payment plan may force card holders to address the reality of their borrowing and help to alter their borrowing behavior or their payoff plans.

REGULATE LATE FEES

One problem with the pricing of credit cards is that credit card firms can charge late and over-limit fees with relative impunity because consumers typically do not believe ex ante that they will pay such fees. Instead, consumers shop based on other factors, such as annual fees, interest rates, or various reward programs. In principle, firms need to charge late and over-limit fees in order to incentivize customers to avoid late fees and going over their credit limits. In practice, given the high fees they charge, credit card firms are perfectly content to let consumers pay late and exceed their limits.

In earlier work, we proposed changing the scoring of the game (corresponding to a regulatory choice in the bottom right of table 26.3). Under our proposal, firms could deter consumers from paying late or going over their credit card limits with whatever fees they deemed appropriate, but the bulk of such fees would be placed in a public trust to be used for financial education and assistance to troubled borrowers. Firms would retain a fixed percentage of the fees to pay for their actual costs incurred from late payments or over-limit charges, and for any increased risks of default that such behavior presages. The benefit of such an approach is that it permits firms to deter "bad conduct" by consumers who pay late or go over credit limits but prevents firms from profiting from consumers' predictable misforecasts regarding their own late payment and over-the-limit behaviors. Firms' incentives to encourage or overcharge for such behaviors would be removed, while their incentives to deter consumer failures appropriately and cover a firm's costs when they occur would be maintained.

ADVANCES IN THE CARD ACT OF 2009

The CARD Act of 2009 enacted a number of key changes to the credit card market that take seriously the behavioral insights and the incentives of firms to exploit consumer failings. For example, the CARD Act provides for improvements in plain language disclosures and timing on credit card agreements. It requires credit card companies to notify consumers forty-five days in advance of certain major changes to card terms, such as interest rates and fees, and it requires that disclosures include information on the time and cost of making only the minimum payment, as well as the time and cost of paying off the balance within three years. Moreover, consumers are provided with monthly and year-to-date figures on interest costs and fees incurred, so that they can more readily compare anticipated costs with their actual usage patterns. The act requires firms to obtain consumers' consent—an opt-in—for over-limit transactions. The act bans practices such as certain retroactive rate hikes on existing balances, late fee traps (including mid-day due times, due dates less than twenty-one days after the time of mailing statements, and moving due dates around each month), and double cycle billing. These practices have in common that consumers cannot readily shape their behavior to avoid the charges; the fees or changes in question are not readily shopped for in choosing a credit card, and disclosures are of little help. Since consumers generally do not understand how payments are allocated across account balances even after improved disclosures (Federal Reserve 2007a,b, 2008), the act requires a consumer's

payments above the minimum required to be applied first toward higher-cost balances. In addition, the act takes up the concern with late fees but goes beyond our proposals. Instead, recognizing that consumers do not shop for penalty fees and that they often misforecast their own behavior, it requires that late fees and other penalty fees be "reasonable and proportionate," as determined by implementing rules; that in any event the fees not be larger than the amount charged that is over the limit or late; and that a late fee or other penalty fee cannot be assessed more than once for the same transaction or event. Furthermore, the act takes steps to make it easier for the market to develop mechanisms for consumer comparison shopping by requiring the public posting to the Federal Reserve of credit card contracts in machine-readable formats. Private firms or nonprofits can then develop tools for experts and consumers to use to evaluate these various contracts. The CFPB will undoubtedly have occasion to review these and other requirements in the future.

Increasing Saving among Low- and Moderate-Income Households

We have focused in this chapter on improving outcomes in the credit markets using insights from behavioral economics and industrial organization. Our focus derives from the relative lack of attention to this area in the behavioral literature thus far and from the fact that credit markets pose a challenge to approaches that do not pay sufficient heed to the incentives firms have to exploit consumer biases. Savings is another area ripe for further examination. Whereas much of the behaviorally informed policy work on saving has thus far focused on using defaults to improve retirement saving, many low- and moderate-income (LMI) households have a much greater need to focus on basic banking services and short-term savings options, services which, for this population, are likely to require a different mix of governmental responses than those in the context of retirement savings for middle- and upper-income households.

Many LMI individuals lack access to financial services, such as checking accounts or easily utilized savings opportunities, that middle-income families take for granted. High-cost financial services, barriers to savings, lack of insurance, and credit constraints increase the economic challenges faced by LMI families. In the short run, it is often hard for these families to deal with fluctuations in income that occur because of job changes, instability in hours worked, medical emergencies, changes in family composition, or a myriad of other factors that cause abrupt changes in economic inflows and outflows. At low income levels, small income fluctuations may create serious problems in paying rent, utilities, or other bills. Moreover, the high costs and low utility of financial services used by many low-income households extract a daily toll on take-home pay. Limited access to mainstream financial services reduces ready opportunities to save and limits families' ability to build assets and save for the future.

In theory, opt-out policies ought to work well among LMI households, as in the retirement world, in encouraging saving. However, while in general the market pulls in the same direction as policy in encouraging saving, market forces weaken or break down entirely with respect to encouraging LMI households' saving. This is simply because the administrative costs of collecting small-value deposits are high in relation to the banks' potential earnings on the relatively small amounts saved, unless the bank can charge high fees; and with sufficiently high fees, it is not clear that utilizing a bank account makes economic sense for LMI households. Indeed, the current structure of bank accounts is one of the primary reasons why LMI households do not have them. High minimum balance requirements, high fees for overdraft protection or bounced checks, and delays in check clearance, dissuade LMI households from opening or retaining bank transaction accounts. Moreover, banks use the private ChexSystem to screen out households who have had difficulty with accounts in the past. Behaviorally insightful tweaks, while helpful, are unlikely to suffice in this context; rather, along with the behavior of consumers who open and maintain them, we need to change the nature of the accounts being offered.

Proposals in this area pertain to changing the rules and the scoring on the left-hand side of table 26.3, where markets may prove neutral to, or even positively inclined toward, the potential reduction of consumer fallibility. We need to figure out how to increase scale and to offset costs for the private sector to increase saving by LMI families. We propose three options: a new "gold seal" for financial institutions in return for offering safe and affordable bank accounts; various forms of tax credits, subsidies, or innovation prizes; and a proposal under which the Treasury would direct deposit tax refunds into opt-out bank accounts automatically set up at tax time. The proposals are designed to induce the private sector to change their account offerings by offering government inducement to reach scale, as well as to alter consumer behavior through the structure of the accounts offered. In particular, the government seal of approval, tax credit or subsidy, or bundling through the direct deposit of tax refunds changes the scoring to firms for offering such products, while the opt-out nature of the proposal and other behavioral tweaks change the starting rules.

One relatively "light touch" approach to improving outcomes in this area would be to offer a government "gold seal" for financial institutions offering safe and affordable bank accounts. While the gold seal would not change the costs of the accounts themselves, it might increase the potency of the bank's marketing and thus reduce acquisition costs; also, the goodwill generated might improve the bank's image overall and thus contribute to profitability. Similarly, small prizes for innovation in serving LMI customers might heighten attention to the issue and increase investment in research and development of technology to serve the poor. Grants to local communities and nonprofits may increase their outreach and improve the provision of financial education and information and help banks and credit unions reach out to LMI customers.

To overcome the problem of the high fixed costs of offering sensible transaction accounts to low-income individuals with low savings levels, Congress could enact a tax credit for financial institutions that offer safe and affordable bank accounts to LMI households (Barr 2004, 2007). The tax credit would be pay-for-performance, with financial institutions able to claim tax credits for a fixed amount per account opened by LMI households. The accounts eligible for tax credit could be structured and priced by the private sector according to essential terms required by regulation. For example, costly and inefficient checking accounts with a high risk of overdraft would be eschewed in favor of low-cost, low-risk accounts with only debit-card access. The accounts would be debit-card based, with no check-writing capability, no overdrafts permitted, and no ChexSystems rejections for past account failures in the absence of fraud or other meaningful abuse.

Direct-deposit tax refund accounts could be used to encourage savings and expanded access to banking services, while reducing reliance on costly refund-anticipation loans and check-cashing services (Barr 2004, 2007). Under the plan, unbanked low-income households who file their tax returns would have their tax refunds directly deposited into a new account. Direct deposit is significantly cheaper and faster than paper checks, both for the government and for taxpayers. Taxpayers could choose to opt out of the system if they did not want to directly deposit their refund, but the expectation is that the accounts would be widely accepted since they would significantly reduce the costs and expedite the timing of receiving one's tax refund. By using an opt-out strategy and reaching households at tax time, this approach could help to overcome the tendency to procrastinate in setting up accounts. By reducing the time it takes to receive a refund and permitting a portion of the funds to be used to pay for tax preparation, setting up such accounts could help to reduce the incentives to take out costly refund loans, incentives that are magnified by temporal myopia and misunderstanding regarding the costs of credit. Such accounts would also eliminate the need to use costly check-cashing services for one's tax refund check. Moreover, the account could continue to be used past tax time. Households could use the account like any other bank account—to receive their income, save, pay bills, and, of course, to receive their refund in following years. There are a variety of ways to structure these accounts, all of which would deploy opt-out strategies and government bundling to reach scale and better align the costs of overcoming consumer bias with the shared benefit of moving households into the banking system. Such an approach could efficiently bring millions of households into the banking system.

The power of these initiatives could be significantly increased if it were coupled with a series of behaviorally informed efforts to improve the take-up of the accounts and the savings outcomes for account holders. For example, banks could encourage employers to endorse direct deposit and automatic savings plans to set up default rules that would increase savings outcomes. With an automatic savings plan, accounts could be structured so that holders could designate a portion of their paycheck to be deposited into a savings "pocket"; the savings feature would rely on the precommitment device of automatic savings, and the funds would be somewhat more difficult to access than those in the regular bank account to make the commitment more likely to stick. To provide the necessary access to emergency funds in a more cost-effective manner than is usually available to LMI households, the bank account could also include a six-month consumer loan with direct deposit and direct debit, using relationship banking and automated payment systems to provide an alternative to costly payday loans. With direct deposit of income and direct debit of interest and principal due, the loan should be relatively low-risk and costless for the bank to service. With a longer payment period than in typical payday lending, the loan should be more manageable for consumers living paycheck to paycheck and would likely lead to less repeated borrowing undertaken to stay current on past loans. Moreover, the loan repayment features could also include a provision that consumers "pay themselves first," by including a savings deposit to their account with every payment. Such a precommitment device could overcome the bias to procrastinate in savings and reduce the likelihood of needing future emergency borrowing. All these efforts could increase take up of the banking product and lead to improved savings outcomes.

The federal government under President Obama has made some progress toward these objectives over the last couple of years. The Treasury Department has launched pilot programs to test different product attributes, including debit cards and payroll cards, and the FDIC has launched a pilot with a group of banks to test consumer demand and sustainability of a safe and affordable account, using an FDIC template, or gold seal. Finally, the Treasury obtained authorization in the Dodd-Frank Act to experiment with a variety of methods to increase access to bank accounts for low-income households, including the provision of seed money for research and development into innovative technology and services.

Conclusion

We have proposed a conceptual framework for behaviorally informed regulation. The framework relies on a more nuanced understanding of human behavior than is found in the classical rational actor model, which underlies much policy thinking. Whereas the classical perspective generally assumes people know what is important and knowable, that they plan with insight and patience, and that they carry out their plans with wisdom and self-control, the central gist of the behavioral perspective is that people often fail to know and understand things that matter; that they misperceive, misallocate, mispredict, and fail to carry out their intended plans; that the context in which they function has great impact; and that institutions shape defaults, planning, and behavior itself. Behaviorally informed regulation is cognizant of the importance of framing and defaults, of the gap between information and understanding and between intention and action, and of the role of decisional conflict and other psychological factors that affect how people behave. At the same time, we argue, behaviorally informed regulation needs to take into account not only behavioral insights about individuals but also economic insights about markets.

In this framework, successful regulation requires integrating a more nuanced view of human behavior with an understanding of markets. Markets have been shown to systematically favor overcoming behavioral biases in some contexts and to systematically favor exploiting those biases in other contexts. A central illustration of this distinction is the contrast between the market for saving and that for borrowing—in which the same fundamental human tendency to underappreciate the impact of compounding interest leads to opposite market reactions. In the savings context, firms seek to overcome the bias; in the credit context, they seek to exploit it. Our framework largely retains the classical perspective of consumers interacting with firms in competitive markets. The difference is that consumers are now understood to be fallible in systematic and important ways, and firms are seen to have incentives to overcome or to exploit these shortcomings.

More generally, firms not only will operate on the contour defined by human psychology but also will respond strategically to regulations. And firms get to act last. Because the firm has a great deal of latitude in issue framing, product design, and so on, they have the capacity to affect consumer behavior and in so doing to circumvent or pervert regulatory constraints. Ironically, firms' capacity to do so is *enhanced* by their interaction with "behavioral" consumers (as opposed to the hypothetically rational consumers of neoclassical economic theory), since so many of the things a regulator would find hard or undesirable to control (e.g. frames, design nuance, complexity) can be used to influence consumers' behavior greatly. The challenge of behaviorally informed regulation, therefore, is to envision not only the role of human behavior, but also the ways in which firms are likely to respond to consumer behavior and to the structure of regulation.

We have developed a model in which outcomes are an equilibrium interaction between individuals with specific psychologies and firms that respond to those psychologies within specific markets. These outcomes may not be socially optimal. To the extent that the interaction produces real harm, regulation could address the potential social welfare implications of this equilibrium. Taking both individual psychology and industrial organization seriously suggests the need for policy makers to consider a range of market-context-specific policy options, including both changing the "rules" of the game, as well as changing its "scoring." We have explored some specific applications of this conceptual framework for financial services.

Notes

1. In addition to incentives to increase savings, employers also seek to boost employee retention, and they must comply with federal pension rules designed to ensure that the plans are not "top heavy." Moreover, there are significant compliance issues regarding pensions and retirement plans, disclosure failures, fee churning and complicated and costly fee structures, and conflicts of interest in plan management, as well as problems with encouraging employers to sign up low-wage workers for retirement plans. Yet, as a comparative matter, market incentives to overcome psychological biases in order to encourage saving are more aligned with optimal social policy than are market incentives to exacerbate psychological biases to encourage borrowing.

2. We use this bimodal framework of regulatory choice to simplify the exploration of how our model of individual psychology and firm incentives affects regulation. We acknowledge that the regulatory choice matrix is more complex (see Barr, 2005).

3. This is largely because of the existing regulatory framework: pension regulation gives employers incentives to enroll lower-income individuals in 401(k) programs. Absent these, it is likely that firms would be happy to discourage enrollment since they often must pay the match for these individuals. This point is interesting because it suggests that even defaults in savings only work because some other regulation "changed the scoring" of the game.

4. This example abstracts from collection costs (which would reduce firms' incentives to hide borrowing costs) and instead focuses on the short-term behavior generally exhibited by firms, as in the recent home mortgage crisis.

5. In the interests of full disclosure, one of us (Barr), was the assistant secretary of the treasury for financial institutions from 2009 to 2010 and led the effort to put in place a number of these reforms in the CARD Act, the Dodd-Frank Act, and other Treasury initiatives.

6. See Federal Reserve Board, Final Rule Amending Regulation Z, 12 CFR Part 226 (July 14, 2008); *Summary of findings: Consumer testing of mortgage broker disclosures.* Submitted to the Board of Governors of the Federal Reserve System, July 10, 2008 (Retrieved from http://www.federalreserve.gov/newsevents/press/bcreg/20080714regzconstest.pdf); Federal Reserve Board, Proposed Rule Amending Regulation Z, 72 Fed. Reg.114: 32948 (codified at 12 CFR Part 226 [June 14, 2007]); Federal Reserve Board (2007a).

7. Although the financial industry often calls for "principles based" approaches to regulation, in the course of the Dodd-Frank Act legislative debate, the industry strongly resisted this approach, perhaps for these reasons.

8. Ian Ayres recently suggested to us that the burden might be placed on the plaintiff to use consumer survey data to show that the disclosure was unreasonable, similar to the process used under the Lanham Act for false advertising claims. In individual cases, this might be infeasible, but such an approach might work either for class actions or for claims brought by the CFPB.

9. Again, in the interest of full disclosure, this proposal was included in the Treasury Department's legislation for the new CFPB but was not included in the final legislation as enacted.

10. See the discussion above relating to the reasonableness standard for disclosure. As noted above, consumer survey evidence could be introduced, either by the CFPB, plaintiffs, or defendants, as to the reasonableness standard.

11. See Jackson and Burlingame (2007). While in principle YSPs could permit lenders legitimately to pass on the cost of a mortgage broker fee to a cash-strapped borrower in the form of a higher interest rate rather than in the form of a cash payment, the evidence suggests that YSPs are in fact used to compensate brokers for getting borrowers to accept higher interest rates, prepayment penalties, and other loan terms.

12. See also Guttentag (2000).

13. See Bankruptcy Abuse Prevention and Consumer Protection Act of 2005, Pub L. No. 109-8, 119 Stat. 23 (codified at 11 U.S.C. § 101 et seq (2005)).

14. See, e.g., U. S. General Accounting Office (2006).

15. See, e.g., Office of the Comptroller of the Currency (2003, 2004a, 2004b).

16. See Federal Reserve Board (2007b).

17. Federal Reserve Board, Proposed Rule, 12 C.F.R. 226, proposed §.7(b)(12), implementing 15 U.S.C. §1637(b)(11).

18. Buehler, Griffin, and Ross (2002); Koehler and Poon, (2005).

19. Barr (2007). For a related proposal, see Gordon and Douglas (2005), in which they argue for an opt-out direct-debit arrangement for credit cards.

References

Adkins, N. R., and Ozanne, J. L. (2005). The low literate consumer. *Journal of Consumer Research*, *32*, 93–105.

Alexander, G. J., Jones, J. D., and Nigro, P. J. (1998). Mutual fund shareholders: Characteristics, investor knowledge and sources of information. *Financial Services Review*, *7*, 301–316.

American Payroll Association. (2002). *Survey results: American Payroll Association 2003 Direct Deposit Survey.* Retrieved from http://legacy.americanpayroll.org/pdfs/paycard/DDsurv_results0212.pdf

Arkes, H. R., and Blumer, C. (1985). The psychology of sunk cost. *Organizational Behavior and Human Decision Processes*, *35*,124–140.

Ausubel, L. M. (1991). The failure of competition in the credit card market. *American Economic Review*, *81*, 50–81.

Bar-Gill, O. (2004). Seduction by plastic. *Northwestern University Law Review*, *98*(4), 1373–1434.

Barr, M. S. (2004). Banking the poor. *Yale Journal on Regulation*, *21*(1), 121–237.

———. (2005). Modes of credit market regulation. In N. Retsinas and E. Belsky (Eds.), *Building assets, Building credit* (pp. 206–236). Washington, DC: Brookings Institution Press.

———. (2007). An inclusive, progressive national savings and financial services policy. *Harvard Law and Policy Review*, *1*(1), 161–184.

Barr, M. S., Mullainathan, S., and Shafir, E. (2008a). *Behaviorally informed financial services regulation.* White paper. Washington: New America Foundation.

———. (2008b). Behaviorally informed home mortgage credit regulation. In N. Retsinas and E. Belsky (Eds.), *Borrowing to live: Consumer and mortgage credit revisited* (pp. 170–202). Washington, DC: The Brookings Institution.

———. (2008c). *An opt-out home mortgage system.* Hamilton Project Discussion Paper 2008-13. Washington, DC: Brookings Institution.

———. (2009). The case for behaviorally informed regulation. In D. Moss and J. Cisternino (Eds.), *New perspectives on regulation* (pp. 25–62). Cambridge, MA: The Tobin Project.

Bell, D. E. (1982). Regret in decision making under uncertainty. *Operations Research*, *30*, 961–981.

Benartzi, S., and Thaler, R. H. (2004). Save more tomorrow: Using behavioral economics to increase employee saving. *Journal of Political Economy*, *112* (1–2), 164–187.

Berry, C. (2004). *To bank or not to bank? A survey of low-income households.* Working Paper Series. Cambridge, MA: Joint Center for Housing Studies.

Bertrand, M., Karlan, D., Mullainathan, S., Shafir, E., and Zinman, J. (2010). What's advertising content worth? Evidence from a consumer credit marketing field experiment. *Quarterly Journal of Economics*, *125*(1), 263–305.

Bucks, B. K., Kennickell, A. B., and Moore, K. B. (2006). Recent changes in U.S. family finances: Evidence from the 2001 and 2004 Survey of Consumer Finances. Federal Reserve Bulletin, *92*. Retrieved from http://www.federalreserve.gov/pubs/bulletin/2006/financesurvey.pdf

Buehler, R., Griffin, D., and Ross, M. (1994). Exploring the "planning fallacy": Why people underestimate their task completion times. *Journal of Personality and Social Psychology*, *67*, 366–381.

———. (2002). Inside the planning fallacy: The causes and consequences of optimistic time predictions. In T. Gilovich, D. Griffin, and D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 250–270). Cambridge: Cambridge University Press.

Cain, D. M., G. Loewenstein, and D. A. Moore. (2005). The dirt on coming clean: Perverse effects of disclosing conflicts of interest. *Journal of Legal Studies*, *34*(1), 1–25.

Cialdini, R. B., Cacioppo, J. T., Bassett, R., and Miller, J. A. (1978). Low-ball procedure for producing compliance: Commitment then cost. *Journal of Personality and Social Psychology*, *36*, 463–476.

Darley, J. M., and Batson, C. D. (1973). From Jerusalem to Jericho: A study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology*, *27*, 100–108.

Edin, K., and Lein, L. (1997). Work, welfare, and single mothers' economic survival strategies. *American Sociological Review*, *62*(2), 253–266.

Engel, K., and McCoy, P. (2007). Turning a blind eye: Wall Street finance of predatory lending. *Fordham Law Review*, 75, 2039–2103.

Federal Reserve Board. (2007a). *Design and testing of effective truth in lending disclosures.* Retrieved from http://www.federalreserve.gov/dcca/regulationz/20070523/Execsummary.pdf

———. (2007b). *Proposed amendments to Regulation Z* (press release). Retrieved from http://www.federalreserve.gov/BoardDocs/Press/bcreg/2007/20070523/default.htm

———. (2008). *Design and testing of effective truth in lending disclosures: Findings from experimental study.* Retrieved from http://www.federalreserve.gov/newsevents/press/bcreg/bcreg20081218a8.pdf

Federal Trade Commission. (2007). *Improving consumer mortgage disclosures: An empirical assessment of current and prototype disclosure forms.* Bureau of Economics Staff Report, Federal Trade Commission, Washington, DC. Retrieved from http://www.ftc.gov/os/007/06/PO25505mortgagedisclosurereport.pdf

Freedman, J. L., and Fraser, S. C. (1966). Compliance without pressure: The foot-in-the-door technique. *Journal of Personality and Social Psychology*, *4*, 195–203.

Gordon, R., and Douglas, D. (2005, December). Taking charge. *Washington Monthly.* Retrieved from http://www.washingtonmonthly.com/features/2005/0512.gordon.html

Gourville J. T., and Soman, D. (1998). Payment depreciation: The behavioral effects of temporally separating payments from consumption. *Journal of Consumer Research*, *25*, 160–174.

Guttentag, J. (2000). *Another view of predatory lending.* Working Paper 01-23-B. Wharton Financial Institutions Center, Philadelphia, PA. Retrieved from http://fic.wharton.upenn.edu/fic/papers/01/0123.pdf

Iyengar, S. S., Jiang, W., and Huberman, G. (2004). How much choice is too much: Determinants of individual contributions in 401(k) retirement plans. In O. S. Mitchell and S. Utkus (Eds.) *Pension design and structure: New lessons from behavioral finance* (pp. 83–97). Oxford: Oxford University Press.

Iyengar, S. S., and Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, *79*, 995–1006.

Jackson, H. E., and Burlingame, L. (2007). Kickbacks or compensation: The case of yield spread premiums. *Stanford Journal of Law, Business and Finance*, *12*, 289–361.

Johnson, E. J., and Goldstein, D. (2003). Do defaults save lives? *Science*, *302*, 1338–1339.

Johnson, E. J., Hershey, J., Meszaros, J., and Kunreuther, H. (1993). Framing, probability distortions, and insurance decisions. *Journal of Risk and Uncertainty*, *7*, 35–51.

Kahneman, D. (1994). New challenges to the rationality assumption. *Journal of Institutional and Theoretical Economics*, *150*, 18–36.

Kahneman, D., and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–291.

Kennedy, D. (2005). Cost-benefit analysis of debtor protection rules in subprime market default situations. In N. Retsinas and E. Belsky (Eds.), *Building assets, building credit* (pp. 266–282). Washington, DC: Brookings Institution Press.

Knetsch, J. L. (1989). The endowment effect and evidence of nonreversible indifference curves. *American Economic Review*, *79*, 1277–1284.

Koehler, D. J., and C.S.K. Poon. (2005). Self-predictions overweight strength of current intentions. *Journal of Experimental Social Psychology*, *42*(4), 517–524.

Koide, M. (2007, November 16). *The assets and transaction account.* New America Foundation. Retrieved from http://newamerica.net/publications/policy/assets_and_transaction_account

Lepper, M. R., Greene, D., and Nisbett, R. E. (1973). Undermining children's intrinsic interest with extrinsic reward: A test of the "overjustification" hypothesis. *Journal of Personality and Social Psychology*, *28*, 129–137.

Lewin, K. (1951). Intention, will and need. In D. Rapaport (Ed.), *Organization and pathology of thought: Selected sources.* (pp. 95–153). New York: Columbia University Press.

Lichtenstein, S., and Slovic, P. (Eds.) (2006). *The construction of preference.* Cambridge: Cambridge University Press.

Loewenstein, G., and Elster, J. (Eds.) (1992). *Choice over time.* New York: Russell Sage Foundation.

Loewenstein, G. and Thaler, R. H. (1992). Intertemporal choice. In R. H. Thaler (Ed.), *The winner's curse: Paradoxes and anomalies of economic life.* New York: Free Press.

Lusardi, A., Mitchell, O., and Curto, V. (2009). *Financial literacy and financial sophistication among older Americans.* NBER Working Paper No. 15469. National Bureau of Economic Research.

Madrian, B. C., and Shea, D. F. (2001). The power of suggestion: Inertia in 401(k) participation and savings behavior. *Quarterly Journal of Economics*, *116*(4), 1149–1187.

Mann, R. (2006). *Charging ahead: The growth and regulation of payment card markets.* Cambridge: Cambridge University Press.

———. (2007). Bankruptcy reform and the sweat box of credit card debt. *University of Illinois Law Review*, *2007*(1), 375–403.

Milgram, S. (1974). *Obedience to authority.* New York: Harper and Row.

Mullainathan, S. and Shafir, E. (2009). Savings policy and decision-making in low-income households. In M. Barr and R. Blank (Eds.), *Insufficient funds: Savings, assets, credit and banking among low-income households* (pp. 121–145). New York: Russell Sage Foundation.

Musto, D. K. (2007). *Victimizing the borrowers: Predatory lending's role in the subprime mortgage crisis.* Working paper, Wharton School, University of Pennsylvania. Retrieved from http://knowledge.wharton.upenn.edu/article.cfm?articleid=1901

Office of the Comptroller of the Currency. (2003). *Account management and loss allowance guidance.* OCC Bull. 2003-1. Retrieved from http://www.occ.gov/news-issuances/bulletins/2003/bulletin-2003-1.html

———. (2004a). *Secured credit cards.* OCC Advisory Letter 2004-4. Retrieved from http://www.occ.gov/static/news-issuances/memos-advisory-letters/2004/advisory-letter-2004-4.pdf

———. (2004 b). *Credit card practices.* OCC Advisory Letter 2004-10. Retrieved from http://www.occ.gov/static/news-issuances/memos-advisory-letters/2004/advisory-letter-2004-10.pdf

Redelmeier, D., and Shafir, E. (1995). Medical decision making in situations that offer multiple alternatives. *Journal of the American Medical Association*, *273*(4), 302–305.

Samuelson, W., and Zeckhauser, R. J. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, *1*, 7–59.

Schultz, E. (1995, December 22). Helpful or confusing? Fund choices multiply for many retirement plans. *Wall Street Journal*, pp. C1, C25.

Shafir, E., Simonson, I., and Tversky, A. (1993). Reason-based choice. *Cognition*, *49*, 11–36.

Sherraden, M., and Barr, M. (2005). Institutions and inclusion in savings policy. In N. Retsinas and E. Belsky (Eds.), *Building assets, building credit* (pp. 286–315). Washington, DC: Brookings Institution Press.

Thaler, R. H. (1985). Mental accounting and consumer choice. *Marketing Science*, *4*, 199–214.

———. (1992). *The winner's curse: Paradoxes and anomalies of economic life.* New York: Free Press.

———. (1999). Mental accounting matters. *Journal of Behavioral Decision Making*, *12*(3), 183–206.

Thaler, R. H., and Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness.* New Haven, CT: Yale University Press.

Tufano, P. (2009). Consumer finance. *Annual Review of Financial Economics*, 1, 227–247.

Tversky, A., and Kahneman, D. (1991). Loss aversion in riskless choice: A reference dependent model. *Quarterly Journal of Economics*, *106*, 1039–1061.

Tversky, A., and Shafir, E. (1992). Choice under conflict: The dynamics of deferred decision. *Psychological Science*, *3*, 358–361.

U. S. General Accounting Office. (2006). *Credit cards: Increased complexity in rates and fees heightens the need for more effective disclosures to consumers.* Report 06-929. Retrieved from http://www.gao.gov/new.items/d06929.pdf

Warren, E. (2007). Unsafe at any rate. *DEMOCRACY: A Journal of Ideas, 5.* Retrieved from http://www.democracyjournal.org/5/6528.php

White, J. J., and Summers, R. S. (1995). *Uniform commercial code*, 4th ed. Practitioner Treatise Series. St. Paul, MN: West Publishing.

CHAPTER 27

Psychology and Economic Policy

WILLIAM J. CONGDON

As this volume amply demonstrates, insights from psychology can and do inform multiple spheres of public policy. From labor law to food and nutrition policy to criminal justice procedures, the role and design of policy, as well as its ultimate effectiveness, depend on how the targeted or affected individuals behave. By offering a scientific, empirically based way of better understanding how humans think, decide, and act, psychological research holds great potential for improving the analysis and design of public policy.

Nowhere is this more true than for economic policy. Already, psychology, under the rubric of behavioral economics, has demonstrated a tremendous promise for informing economic policy. Most famously, insights from psychology have had a clear impact on retirement savings policy, as touched on in numerous chapters in this volume. The practice of automatic enrollment in 401(k)s has been codified in the Pension Protection Act of 2006. Another recent reform makes it easier for individuals to direct a portion of their income tax refunds to retirement savings accounts. And current proposals include the establishment of more automatic forms of IRAs and the simplification of retirement savings tax credits available to low- and moderate-income families.

Each of these policy innovations has a direct antecedent in research in behavioral economics. The dramatic success of automatic enrollment in 401(k) plans provides the most striking example. In one influential study, participation rates among new workers in one company's 401(k) plan went from 37% to 86% after switching from a traditional enrollment scheme to one in which workers were automatically enrolled but could later opt out (Madrian and Shea, 2001). Some estimates suggest that the total increase in 401(k) balances due to the move to automatic enrollment may be as much as $300 billion over ten years (Iwry, Gale, and Orszag, 2006).

The excitement and the resulting policy reform generated by the automatic enrollment findings have been due to a number of factors. From the perspective of policy, this was a highly desirable result. The policy goal of encouraging retirement savings is a long-standing one. Automatic enrollment was demonstrably effective. Moreover, as policies to promote retirement savings go, automatic enrollment has been cheap. It leads individuals to contribute more to their 401(k) plans without requiring, for example, the government to increase the tax subsidy.

From the perspective of behavioral economics and from that of the larger project of informing economic policy with insights from psychology, these results and the subsequent policy showed the real-world relevance of behavioral economics. Any lingering doubt that behavioral economics was a purely academic exercise, collecting fascinating but ultimately inconsequential examples of human foibles, was largely put to rest. The automatic enrollment result was one that standard economic models would not have predicted and a policy reform they did not suggest.

Based in part on this success, policy makers have shown an interest in considering the behavioral dimensions of everything from labor-market policies to health-care reform to environmental regulations. And economists and behavioral science researchers have sought to apply lessons from psychology in many of these areas, as evidenced both by the wide variety of economic policy topics considered in this volume, as well as by other research and writing.

But while behavioral economics is clearly a powerful tool for policy, it is not clear that we know how best to wield it. Should we take defaults as a new lever for economic policy and look for other domains in which to apply them? Or were defaults a lever specific to retirement savings? How do we search for other such levers? Moreover, while defaults achieve some narrow goals associated with participation in retirement savings plans, how do they relate to broader social goals related to life-cycle saving and retirement security? Does the success of defaults in encouraging participation in retirement savings plans raise any larger questions about how policy ought to be structured in that domain?

The chapters in this book related to economic policy document many successes in applying psychological lessons to economic policy, but they also reflect

some of this uncertainty. This uncertainty arises most clearly in the *application* of psychological insights to economic policy questions. Given some set of policy objectives—encouraging retirement savings, discouraging carbon emissions, alleviating poverty, and so on—how can psychological insights improve policy design? Economic analysis is still, to a large extent, in an age of discovery when it comes to learning about the practical applications of psychological concepts and findings to questions of economic policy.

This hesitancy reflects, in part, a deeper uncertainty related to the broader *approach* that economics should take in incorporating psychological insights into policy analysis. Findings from psychology raise difficult questions for policy that economics is not accustomed to tackling—if individuals can make mistakes or procrastinate, how should economic policy deal with or reflect that possibility? Moreover, findings from psychology might not only inspire us to rethink the design of policy responses, but also cause us to rethink the nature of underlying problems. If procrastination is the problem in retirement saving, are 401(k)s the solution?

Building an Approach

In general, even where insights from psychology are relevant to economic policy, they cannot be applied to policy analysis directly. For example, research in psychology has determined that individuals have only limited attention. This is likely to be relevant for policies related to life-cycle saving or curbing carbon emissions, but the application is not mechanical. The finding alone does not imply anything specific for economic policy. Likewise, economic policies rarely beg particular psychological insights. What is needed, then, is an approach for integrating psychological insights into economic policy analysis.

We need to do this for economic policy for at least two reasons. One is to provide a framework for *how* to apply psychological insights. This is the key issue for informing the design of economic policy. Given findings from psychology, how should we think about the implications for the tools that economic policy typically employs? What are the implications of limited attention for, say, the effectiveness of corrective taxes? Moreover, what new design tools does psychology add to our tool kit, and how do they operate? Channel factors might imply the power of, for example, defaults in social programs—and so on.

But we also need an approach in order to provide a filter that tells us *what* to do with those insights. Economic policy analysis is restrictive. It identifies some situations—for example, market failures—as problems economic policy can rectify and leaves others aside. For a behavioral approach to economic policy we have to ask, Do findings from psychology change how we think about the underlying problems we are trying to solve in addition to giving us new tools with which to solve them? Should we take, say, findings on time-inconsistent behavior as an invitation to design corrective taxes to curb activities it may give rise to, such as smoking? Or should we maintain the traditional posture of economics of not interfering with such choices when they create no externality?

Choice Architecture, Nudges, and Asymmetric Paternalism

Much existing work in behavioral economic policy has coalesced around a set of related approaches that go by various names: choice architecture, nudges, libertarian paternalism, asymmetric paternalism, and so on. While distinct in some important ways, they share many common features along two key dimensions: the filter they use in settling on which policy problems to address and the framework they use in determining how to address them.

CHOICE ARCHITECTURE AND NUDGES

In their contribution to this volume, Thaler, Sunstein, and Balz designate those individuals who have a role in setting the context in which others make decisions as *choice architects*. While that term is not specific to economic policy, economic policy makers fit squarely under this heading. Economic policy is in many ways the business of structuring and influencing choices. Whether it is setting rules about choosing between the early and normal retirement age in Social Security, or establishing the exchanges through which individuals will purchase health insurance policies under the new health reform act, or writing regulations intended to curb the activities of individuals that create excess carbon emissions, economic policy shapes the context of choice.

The key conclusion from their work, which is its central psychological insight, is, as they put it, that everything matters. By that they mean that every aspect of the choice context can be expected to exert some influence on how individuals will respond, even the minor or the tangential aspects—and sometimes especially those. The reason that this mantra—everything matters—is a powerful statement for the purposes of economic policy is that in economic policy analysis as it is typically practiced, a lot of things are assumed not to matter. For example, default rules should be relatively neutral in the standard analysis—the costs of, say, filling out a form to enroll in a 401(k) are

so low compared with the potential benefits as to be negligible. Similarly with presentation effects. Or hassle costs. And so on. What psychologists have long known, however, and what behavioral economists appreciate, is that, empirically, these things do matter. And so the question becomes how economics might account for and make use of this fact in analyzing and setting policy.

Perhaps the seminal contribution of this line of research—which is outlined in their chapter and developed in expanded form in Thaler and Sunstein's book, *Nudge* (2008)—is in the way it goes beyond the mere observation that everything matters to give prospective choice architects guidance as to which things matter that they might have previously overlooked. And in doing so, they give a sense of how to make use of these influences to subtly but powerfully reshape choices: how to *nudge*, in their terminology. They present a high-level set of insights into the art and science of nudging: attend to the power of defaults; expect that individuals will err; give feedback to improve choice; assist individuals in understanding mappings from their choices to welfare; structure complex choices; and don't forget incentives.

While this approach is not limited to economic policy, it has informed how we understand the role of psychology in economic policy deeply. The experience with automatic enrollment in retirement plans can easily be viewed through the lens of choice architecture. And the insights here are structured in such a way that they are prescriptive, as well, so that, for example, it is easy to work through the direct application of these insights to a policy challenge, such as setting up the health insurance exchanges that will need to be established in coming years as the health-care reform law takes force. Are the choices among insurance plans well structured? What are the defaults for plan choice and enrollment? Do they help individuals understand the welfare implications of alternative plans?

In terms of informing the approach by which insights from psychology might be brought in to economic policy, we can see that choice architecture is first and foremost a framework for *how* to apply psychological insights to policy questions. And it is a how-to guide par excellence. A vast set of results from psychology and behavioral economics are distilled down to a few key, easy-to-apply points about how policy works and can work when individuals are not as economics assumes them to be but rather how we find them in the world. Thaler and Sunstein even developed a mnemonic, NUDGES, for their core set of insights.

The approach in their chapter with Balz also embeds a filter, a way of dealing with the question of what we should try to do with these insights, in what the authors refer to here and in other work as libertarian paternalism: that these insights can be used to influence individuals to make better choices according to their own judgments without restricting their choices (Thaler and Sunstein, 2003). This idea is closely related to a variant called asymmetric paternalism, which receives a fuller treatment in the chapter in this volume by Loewenstein, John, and Volpp.

ASYMMETRIC PATERNALISM

The chapter by Loewenstein, John, and Volpp gives its own very successful set of advice for incorporating insights from psychology into attempts to address social problems at all levels, including through economic policy. While broadly in line with the choice architecture approach, their particular insight is to note that the very behavioral tendencies that sometimes complicate policy-making *challenges*—procrastination, biases in risk assessment, and so on—also create new *opportunities* for policy making. In one particularly compelling example, they show how an intervention that takes advantage of the very same attractiveness of lotteries that standard economic analysis might conclude to be suboptimal can be used as an effective incentive for encouraging adherence to prescription medication.

This chapter is somewhat more explicit about the filter that is applied in determining *what* social problems this approach addresses, not just how to solve them. The specific approach outlined here is known as asymmetric paternalism (Camerer et al., 2003). The authors note that asymmetric paternalism has two essential features. First, it recognizes that in a world where everything matters, paternalism is unavoidable. Defaults are a classic case. In something like a retirement savings plan, there has to be a default, one way or the other. In the standard economic model, setting the default is of little significance because individuals will not be strongly influenced by it—the cost of filling out a form one way or the other is trivial. And so setting the default one way or the other is not fraught with paternalistic implications. In a behavioral world, the default has substantial consequences—we know that individuals tend to stick with defaults. The choice of how to set the default now becomes more loaded; some measure of paternalism becomes unavoidable.

The second, and maybe defining, feature of asymmetric paternalism is that it sets a policy agenda of helping individuals to help themselves in ways that preserve choice. If the challenge created by behavioral tendencies is that paternalism is unavoidable, the opportunity it creates is to influence behavior without restricting choice. Defaults, again, illustrate the point. They influence individuals to choose one way

or the other but do not preclude the possibility that individuals might still opt in or out if they wish. So in a situation such as retirement savings, where there is reason to think that behavioral tendencies such as present bias lead individuals to save too little on their own, switching the defaults in retirement plans from opt-in to opt-out might help people to help themselves save more without forcing them to follow any particular course of action.

What this example highlights is how this work arrives at a key set of insights for folding psychological findings into economic policy in terms of determining *what* to do, not just how to do it. What we might do, in this approach, is improve choice. Behavioral tendencies can make it hard for people to know best what to do or can impair their capacity to realize their desires. Policy will rarely be neutral in this regard—it will invariably either make it easier for people to make good choices, or not; and we should try to help people make good choices. Policy should help individuals make, for example, patient choices in retirement savings by encouraging participation in employer plans. And, by extension, we could look elsewhere and, say, improve choice in picking the optimal retirement age in Social Security, or the best health insurance plan, or the right car—where optimal, best, and right are always as judged by the individuals themselves.

In addition, what this approach does with great elegance is to deal with the issue of how policy should seek to influence choice where doing so is inherently paternalistic. Economics as an approach to policy is somewhat uncomfortable with saying which choices are better or worse for individuals. Who is to say which choices are a result of procrastination or error, as opposed to simply unusual preferences or circumstances? That individuals should be saving more for retirement? Choose a different health insurance plan? Buy a different car? And the answer this approach gives is to say, we'll nudge people, not shove them. We will structure choices in ways that we think will lead them to decisions they will ultimately be happier with, but we will leave open the possibility that we are wrong for any particular individual and preserve their ability to do otherwise if they so choose.

BENEFITS AND LIMITATIONS

The general approach reflected in these works is, if not the consensus approach in economic policy of incorporating insights from psychology, in some ways the currently ascendant one. This is for good reason: it works. No higher compliment can be paid. Defaults in 401(k) plans increase plan participation. Lotteries improve adherence to medication regimens—and so on. This approach is grounded firmly in both research in psychology and behavioral economics, as well as in institutional knowledge of social problems and how economic policy operates. It is above all empirical and scientific.

The chapter in this volume by Johnson and Goldstein illustrates these points by focusing in on the case of the D in NUDGES—defaults. By providing numerous examples—from auto insurance to organ donation—they show how defaults have consistent power across a wide set of domains and in a variety of particular forms and so represent an effective, and often efficient, lever for policy makers. In addition, they show how the power of defaults is grounded in specific features of psychology, such as the role of implied endorsement and loss aversion in decision making.

Moreover, this approach provides not only guidance on the matter of where to shine the light of behavioral economics, but also a way of resolving, or at least managing, the legitimately quite tricky welfare problems that behavioral economics can raise. It identifies as targets of policy cases where we can help people help themselves. Behavioral tendencies like procrastination mean that people might, for instance, save too little for retirement; policy can help them save more. And it also resolves the big question that gets raised here, how to make these judgments—are we in a good position to tell people they should save more? And the answer is sort of a hedge: to set policy so as to encourage the one without precluding the other. This is an insightful, artful, and eminently reasonable way to deal with this issue. Above all, it is a profoundly humble approach.

The humility of this approach, as commendable as it is, comes at a cost, however. The chief cost being that, because it integrates psychology into economic policy analysis in this deliberately circumscribed way, it is in some ways a local approach to behavioral economic policy.

We can see this first in the way that this approach informs the matter of *what* behavioral insights should lead us to do. By focusing on the class of problems for which we can help people help themselves, we miss a lot of what is interesting in economic policy. For example, in the case of environmental externalities associated with carbon emissions, the goal is to lower overall levels of emissions. The primary concern is not about helping individuals make better choices about, for example, automobile purchases. The policy goal is about directing and encouraging private behavior in such a way as to correct for the social problem, the externality. And the policy response—for example, a carbon tax—might leave individual consumers worse off. If nudges can be effective levers in addition to or

in place of traditional levers, such as fees and taxes, we may want to use them here, too. And we might want to use them to achieve social ends even when they are not helping individuals privately.

Even where we think that helping people help themselves is not at odds with, or even is in alignment with, broader social goals, casting the problem in this relatively restrictive way might lead to unintended consequences. From the perspective of economic policy, the evaluation of a policy such as automatic enrollment in retirement savings plans cannot stop at the level of seeing increased participation and savings and conclude that this is on net a good policy. Economic policy analysis has to be more global. Does automatic enrollment achieve the broader goal of promoting retirement security? For example, where does the money come from that individuals put into these plans when influenced by defaults? Do individuals reduce consumption correspondingly, do they substitute from other forms of savings, or something else? Too narrow a focus might result in policy recommendations that are locally effective but globally counterproductive.

This approach is also somewhat local in application in the matter of *how* to incorporate psychological insights into policy design—first, in that it can focus on improving how policy parameters are set without sufficiently questioning the implications of behavioral economics for the functional form of policy. So, for instance, psychological insights underpin the use of automatic enrollment in retirement savings plans. This change in the design of these plans surely makes defined contribution plans a more effective vehicle for individuals in terms of accomplishing the goals of life-cycle saving. But those findings also raise larger questions about the relative weight that retirement security policy should place on individual savings in relation to such alternatives as Social Security. In this way, this approach can raise important questions about the nature and structure of policy.

Also in terms of policy design, this approach is somewhat local in the sense that focusing on nudges in isolation can miss how they interact with other economic forces. Working from some policy context as a starting point and seeking to apply nudges runs the risk of missing how behavioral levers might interact with other relevant features of the economic context. For example, defaults are effective at improving program participation, but behavioral levers like defaults might interact with economic forces such as asymmetric information to have undesired results, such as subverting efficient screening into social programs.

To be sure, behavioral policy researchers are not on the whole unaware of these challenges. For example, in their chapter in this volume Miller and Prentice, as part of a broader analysis of the design of psychological policy levers, highlight some specific manifestations of this last point—how psychological levers and economic forces can interact, sometimes in counterintuitive ways. They give examples where, for instance, financial incentives—such as taxes and subsidies—and nudges appear to crowd one another out rather than reinforce each other. What we seek for a behavioral approach to economic policy is an approach that can systematize these sorts of insights across the full range of relevant economic forces.

Behavioral Field Economics

Alternative approaches to incorporating findings from psychology into economic policy analysis are possible and can address some of these limitations, even if they introduce others of their own. One promising approach is to build insights from psychology directly into policy-relevant fields of economic analysis, deriving the policy consequences that result from updating standard assumptions about choice and preferences. Fields of economic analysis, such as macroeconomics, public finance, and so on, provide an existing, developed framework for economic policy analysis. Distinct fields provide their own way of identifying what problems economic policy should solve. They also provide a natural way of thinking about how insights from psychology might inform policy design.

BEHAVIORAL PUBLIC FINANCE

A good example of the potential for deriving policy implications of psychological insights through a behavioral approach to a field of economics is behavioral public finance (Congdon, Kling, and Mullainathan, 2011). Public finance provides a natural field into which we might incorporate insights from psychology and behavioral economics. To begin with, much of economic policy is located here: externalities, social insurance, redistribution, tax policy, and so on. And public finance provides a fairly comprehensive analytical framework for understanding the questions that it considers: from thinking about the nature of market failures and how or when they arise and what their welfare consequences are, to how we think about the trade-offs policy makers face and how we design policy responses.

One way to organize the method of public finance is to think of it as proceeding in three stages: performing a diagnosis of the market failure; defining judgments that policy makers must make in addressing market failures; and offering prescriptions for policy

design in response to those failures. Behavioral economics informs public finance at each of these levels.

Diagnosis

Public finance provides a set of policy problems: markets fail, social welfare might be improved through redistribution, and the operations of the state must be financed. A behavioral approach to public finance considers how these problems might change if individuals are imperfect decision makers. The key point is that deviations from the usual economic assumptions about choice and behavior have spillover effects on the operation of markets and for public policy. That is, nonstandard preference and choice behavior matter not just for individual outcomes but also for collective outcomes. So, for instance, behavioral public finance might start by investigating how behavioral tendencies, such as present bias, mediate the impact of environmental externalities, or how limits to computational capacity interact with adverse selection in health insurance markets, or how limits to attention affect individuals' responses to taxes.

Starting from places where government already has an interest allows us to redirect the focus of behavioral insights for policy away from that narrow class of cases in which policy can help individuals help themselves and toward a larger set of policy questions. For example, in the area of tax policy, starting from psychology and working toward policy results in the idea of sin taxes, which, when levied on goods such as alcohol or cigarettes, can help imperfectly rational individuals to make better choices and improve their welfare. But starting from public finance results in a large set of fundamental questions about tax policy—about how to raise revenue efficiently and equitably. And since questions of tax efficiency and equity depend intimately on the behavioral response to taxes, these are matters of policy that behavioral economics speaks to directly. If individuals fail to respond to a tax because they fail to notice it, what does that mean for the excess burden of the tax? And who bears its incidence?

Thinking about the impact of policies in terms of social welfare in public finance also provides policy analysis with a way to think systematically about the global versus the local impacts of policies. Evaluating policy alternatives within this schema helps to avoid the possibility that applying behavioral insights to policy too narrowly will result in outcomes that appear to be improvements along one dimension but end up ignoring other dimensions along which the policy might represent a deterioration. For example, it provides a way to think about how to incorporate behavioral insights into our understanding of retirement security and old-age insurance more broadly, rather than relying on narrow outcomes related to increasing retirement savings through particular channels.

Judgment

In addition to giving behavioral economic policy scope, public finance can also give it shape. In particular, public finance buys applied behavioral work some purchase on the questions that psychology raises about what individuals really want and whether public policy should steer individuals toward one or another choice that is better in some sense. Who is to say, for example, that individuals should save more for retirement or smoke fewer cigarettes?

Public finance is, in fact, already familiar with incorporating judgments of this kind in the way it treats distributional issues. And what it shows in that treatment is that as a practical matter, economic policy analysis can proceed productively without definitive conclusions on such questions. Public finance leaves the determination of the appropriate welfare weights across individuals—the resolution of interpersonal welfare conflicts—as exogenous to economic analysis. By analogy, it suggests that a similar approach to matters of intrapersonal welfare conflicts—such as those that might arise between say, a short- and a long-run self—might be sufficient, leaving to policy makers and society at large the question of, say, whether or not to encourage retirement saving in individuals who appear to have time-inconsistent preferences. This perspective allows us to move past debates about identifying true utility and focus, as with traditional public finance, on policy design conditional on judgments on such matters.

It is also worth pausing here to note that while this approach, of setting these questions aside, provides a workable way forward for policy analysis, public finance also provides a framework for taking a deeper look at these questions. For public finance as a scientific endeavor, these are central questions, even if policy can proceed without resolving them. But even here, public finance provides a starting point and framework for this analysis. The most complete work to date on this fundamental question is found in a set of papers by Bernheim and Rangel (2007, 2009), who show some of the challenges and difficulties associated with this issue.

Prescription

Finally, by embedding psychology into public finance we can then derive principles about the design of policies that reflect the interaction of behavioral tendencies and economic forces. The first lesson behavioral public finance yields is that prices and incentives, the key policy levers in economics, interact with psychological forces. An example is found in the role

of moral hazard in economic policies. For instance, in the case of unemployment insurance, what looks like moral hazard—say, increased unemployment spell length associated with benefits—might not be a straightforward result of the incentives that the program creates but might instead be due to psychological elements of how people respond to these incentives, such as the effects of time-inconsistent preferences. And this might then inform program design. For example, bonuses for finding a job can theoretically serve to realign incentives and mitigate moral hazard but might not be effective if individuals are present biased.

A second important design lesson from the integration of public finance and behavioral economics is that behavioral tendencies are likely to interact with information asymmetries. For example, in the standard model, adverse selection can cause markets to be fragile or fail outright. Behavioral economics suggests that private information does not necessarily operate as in the standard model. For example, adverse selection in health insurance markets might be mitigated if behavioral tendencies such as biases in risk assessment mean that individuals are unable to perceive or act upon information advantages they have with respect to their own health status. Similarly, many results in traditional public finance depend on prices and incentives serving to screen in an efficient way. But behavioral tendencies might undo or reverse the standard conclusions about which population particular incentives target—so that, for example, policy can no longer assume that the transaction costs associated with taking up social benefits are screening applicants efficiently.

Finally, a third set of design insights that come out of this approach is to see how psychological forces interact with market forces. Markets operate based on the choices individuals make, but when those choices become disconnected from preferences due to, for example, choice errors or failures of self control, the outcomes can be inefficient. One result is that competition can operate along dimensions other than what policy makers expect or intend. As a result, policies that seek to harness market forces have to consider the impact of behavioral tendencies. A recent example of this lesson is the case of the Medicare prescription drug benefit. The benefit was organized as a marketplace where seniors could choose subsidized coverage from private providers. In the rational model, the creation of this marketplace should enhance efficiency. But in practice this market is difficult for individuals to navigate: participants choose from dozens of plans that vary on multiple dimensions. As a result, not only did individuals likely make costly private errors with regard to the Medicare drug benefit, the beneficial effects of market competition likely did not manifest.

OTHER FIELDS

In a similar manner, we can think about folding behavioral insights into other fields of economic analysis that contribute to economic policy. Just as behavioral public finance might work through the role of limited attention or nonstandard preferences for negative externalities and their correction, so too, for example, might behavioral macroeconomics consider the implication of such psychological forces for, say, business cycles and their mitigation (Akerlof and Shiller, 2009). And similarly for many other such fields of economic analysis.

Indeed, a number of the chapters in this volume concerned with economic policy or topics related to economic policy can be thought of as loosely coalescing around, and potentially contributing to, distinct fields of economic inquiry concerned with economic policy. For example, the chapter by Barr, Mullainathan, and Shafir develops a framework for thinking about interactions between markets, firms, and the psychology of decision making and derives implications for regulation. In another chapter, Fischhoff and Eggers discuss, among other issues, how regulatory policy should think specifically about disclosure when the targets are imperfectly rational. These chapters fit naturally with research on behavioral industrial organization, which considers, for example, the role of consumer biases and heuristics in the creation or perpetuation of market power and policy responses (Ellison, 2006).

Still other chapters collect insights that might contribute to a behavioral approach to labor economics. The chapter by Jolls provides insights on, among other topics, the role behavioral forces play in explaining the structure of and responses to labor compensation arrangements, including minimum wage policies. And the chapter by Garcia and Cohen considers behavioral dimensions to educational performance, along with some guidelines for policy interventions. These, too, are of a piece with other research in the more general project of working through the implications of psychology for labor market policy (Babcock et al., 2010).

Implementing Applications

The point in developing abstract approaches to behavioral economic policy is, of course, to apply them to concrete policy questions. With such application, we can see how differences in approach work to generate different perspectives and conclusions. Applications that reflect more of a choice architecture approach tend to find specific modifications to

policies that follow from specific psychological insights. Applications more along the behavioral field approach are often less specific but can be broader. Many applications reflect influences of these alternative approaches without being explicitly attached to a particular school of thought and reflect some elements of either approach.

Retirement Security

The signature policy application of behavioral economics to date involves the topic of retirement security. The government sponsors numerous programs in the name of supporting the consumption of retired workers—including subsidies to retirement savings through tax incentives such as IRAs and 401(k)s and through the old-age insurance component of Social Security—and behavioral economics can potentially inform a number of these at a variety of levels. Automatic enrollment in employer retirement plans, as discussed above, is only the most well known behavioral insight for retirement security. As the chapter in this volume by Benartzi, Peleg, and Thaler outlines, behavioral insights have been used not only to increase participation in retirement savings plans, but also to increase contributions to those plans and to improve diversification of retirement portfolios.

The applications that yield these results are quintessential examples of choice architecture. For example, policies to increase contributions to retirement savings take the form of automatic escalation of contributions in plans: individuals' contributions increase automatically either on a schedule or when they receive a raise. These programs take advantage of specific psychological insights, such as hyperbolic discounting, inertia, and loss aversion, to encourage individuals to save more for retirement. Similarly, the menu and nature of investment options available to individuals through retirement plans can affect how individuals allocate their savings. Given the goal of helping people to help themselves, this approach is extremely successful. There can be little doubt that many individuals are better off for the implementation of these design features.

But these findings also illustrate some of the limits to this approach. Whether retirement security policies are meeting their broader social goals cannot be evaluated solely on the grounds of how effectively they help people to help themselves. Stepping back, it is less obvious that policies like automatic enrollment and escalation are always globally beneficial. For example, where does the extra money of those individuals, who are defaulted into 401(k)s but would not otherwise sign up, come from? Upon becoming enrolled, their paychecks are reduced. How do they respond? Do they buy less stuff? Or do they keep on consuming as before and just run up higher credit card debt? Do they stop contributing to other retirement savings vehicles, like IRAs? Knowing the overall welfare impact of automatic enrollment depends on answers to questions such as these. The question of whether automatic enrollment is a good idea is more difficult than it first appears.

The effectiveness of such design features also raises other, larger, questions about the retirement savings policy and the optimal role of behavioral economics in it. For example, what do the automatic enrollment results say about the desirability of the functional form of the policy, that is, 401(k) and 401(k)-type vehicles, as a way to promote retirement security? If such a minor change in program rules can have such dramatic effects on participation, are tax incentives the right model for encouraging retirement savings? Moreover, is the tax incentive attracting the right people into the program relative to the enrollment process? If the problem with retirement savings is, in part, that individuals will put off things like completing applications, then incentives so remote in time as tax breaks in retirement may not be a good solution to this problem. Given what these findings seem to indicate about the attentiveness and commitment of individuals to saving for their own retirement, what does this say about the desirability of individually directed retirement savings relative to alternative policies, such as, say, add-on private accounts in Social Security?

Environmental Externalities

Another set of economic policy issues with important behavioral dimensions are those related to environmental protection, especially those attendant on the problem of carbon emissions and global warming. Economics already has a well-established framework for assessing the inefficiencies associated with excess carbon emissions, and for designing policy responses, by modeling the problem as a negative externality: the costs that carbon emissions impose on society are outside the calculus of any individual agent, and so activities that lead to those emissions will tend to take place at inefficiently high levels. And following from this analysis, economics offers a menu of policy responses to correct for the externality—for example, imposing carbon taxes or issuing tradable permits.

However, because carbon emissions are ultimately mediated by the psychology of consumers who demand carbon-intensive goods, forms of transportation, and energy supplies, insights from psychology can potentially inform how we think about outcomes in the face of the externality and possible corrective action. The chapter in this volume by Weber gives some applications of psychological insights to this

problem, focusing both on the behavioral tendencies that complicate how individuals process and respond to environmental risks and on the possibility that interventions that operate on those tendencies might encourage environmentally protective behavior.

Environmental externalities are a good example of a policy problem that does not particularly fall under the heading of helping individuals help themselves. To be sure, one of the central insights from a behavioral analysis of this problem is that individuals may sometimes fail to take proenvironmental behaviors that would appear to be to their own benefit, for example, by failing to purchase and install energy-efficient lightbulbs. And there are reasons to suspect that behavioral tendencies play a role in this outcome. Present bias, for instance, might depress investments in energy-saving technologies because of the way in which the costs of such actions are front-loaded, while the benefits are realized only in the future. Compact fluorescent lightbulbs save money down the road, but they cost more than incandescents today.

However, while this behavior certainly complicates the policy problem, the first-order social goal of climate policy is not to leave individuals better off by helping them to make better environmental decisions; it is to reduce carbon emissions to stave off the externalities associated with global warming. And in addition to creating this complicating factor, where policy may have to correct not only for the externality but also for psychological forces working against it, as Weber's chapter outlines, these psychological forces also create opportunities. The opportunity is that policy makers might make use of behavioral levers—nudges—to effect reductions in carbon emissions. For example, social comparison is a promising device for nudging individuals toward lower levels of energy consumption. And other such interventions are possible, such as framing, the use of defaults, and so on. What is clear here, again, is that we are less concerned with whether the choice architecture is helping individuals to achieve their private optimum and more concerned with moving overall consumption toward a social optimum.

While behavioral levers show much promise for addressing environmental externalities, an analysis that integrated psychological insights into standard economic approaches to externality correction could potentially go even further. In particular, rather than looking at behavioral levers as a largely separate menu of policy options, we can consider how they operate jointly with standard measures such as taxes. For example, we might be able to use behavioral insights to make corrective taxes work better. One of the reasons that corrective taxes might fail to bring about desired levels of behavior change is that some forms of taxes might fail to be salient. Policy could make a point of setting taxes in ways that individuals will attend to, such as requiring posted prices to reflect taxes, or levying taxes upstream so that they are reflected in prices faced by consumers. Similarly, taxes might fail as a corrective measure because the price schedules they modify are complex or opaque. Innovations to make the relationship between behavior and cost might lead corrective taxes to be more effective. And this street might go both ways: taxes might also be able to effect nudges. For example, taxes on particular goods might send signals about social approval or disapproval of particular behaviors.

Poverty Alleviation

Finally, consider the policy challenges associated with alleviating poverty. Poverty remains a serious issue in the United States, and a broad portfolio of policies—including traditional cash transfers, tax credits, food and housing assistance, and subsidized health insurance—seek to address the hardships associated with poverty. Behavioral tendencies are certain to mediate both the problem of poverty and the success of efforts to address it, and researchers have recently begun to attempt to understand poverty through a behavioral lens. The chapter in this volume by Mullainathan and Shafir provides a behavioral take on elements of the question of poverty.

The distinctive approach of their chapter is its attempt to understand what behavioral economics implies for how poverty arises and perpetuates, and how it impacts welfare, before proceeding to look for implications of behavioral economics for policy design. The authors emphasize the interaction of the type of financial instability that is endemic in the lives of the poor and behavioral tendencies that might magnify the impact and consequences of that instability. They argue, for instance, that dealing with this instability draws on the limited computational capacity of individuals, leaving them with diminished cognitive resources with which to address other issues. Likewise, they suggest that when worn down by the grind of managing this instability, these individuals will have depleted self-control. This shows some of the promise of this type of deep union of behavioral insights and economic analysis: it can change our understanding of the underlying problem and, in doing so, bring the nature of the policy challenge into better focus. Here, for example, it suggests that antipoverty efforts should target not only income or consumption levels but also their volatility.

This approach generates implications for the design of antipoverty policy. For example, if economic volatility is a central issue in poverty, policies that build buffer stock savings, such as individual development accounts, might be a more important part of the

solution than they are in standard models. Turning to programmatic services, such as subsidized housing or nutrition benefits, this approach leads to proposals to better deliver services, such as through simplified enrollment and eligibility procedures. Here a more completely integrated approach could ultimately be richer still. For example, the economics of benefit programs is interested not just in the delivery of benefits but also in their targeting. Those programs operate most efficiently when the take-up process serves to screen applicants such that those who take up benefits are those who benefit from them the most. Behavioral tendencies are likely to interact with this screening. And different design changes to improve take-up, such as simplifying enrollment versus changing defaults, might interact with these tendencies differentially, so that one or the other might be better or worse for targeting.

And a full behavioral analysis of antipoverty programs could go further. For example, another key challenge in program design is mitigating moral hazard. As discussed above, the incentives that lead to moral hazard might interact with behavioral tendencies so as to change both the nature of the problem as well as effective solutions. Policies may have to be as attuned to their tendency to, say, lead to procrastination or compete for limited attention as to their financial incentives.

Going Forward

Recent years have brought truly exciting developments in economic policy in the form of behavioral economics, and this volume collects many excellent contributions to that effort. Insights from psychology appear to offer great promise for improving both our understanding of the problems we want to solve with economic policy and the effectiveness of those policy responses. The choice architecture model, which is well represented here, has been enormously influential, and with good cause.

But while nudges and asymmetric paternalism have represented a tremendous leap forward for behavioral economic policy, such an approach is not without limitations. One promising alternative approach is to integrate behavioral economics more fully into the policy-relevant fields of economics. A behavioral approach to public finance is one example, and chapters in this volume hint at other fields that are ripe for scrutiny through a behavioral lens, such as labor economics and industrial organization. The impact of behavioral economics for a diverse set of policy challenges, including some of those covered in this volume—retirement security, environmental externalities, and poverty alleviation—can potentially be extended and expanded through such an approach.

References

Akerlof, G. A., and Shiller, R. J. (2009). *Animal spirits: How human psychology drives the economy, and why it matters for global capitalism.* Princeton, NJ: Princeton University Press.

Babcock, L., Congdon, W. J., Katz, L. F., and Mullainathan, S. (2010). *Notes on behavioral economics and labor market policy.* Discussion Paper, Brookings Institution.

Bernheim, B. D., and Rangel, A. (2007). Behavioral public economics: Welfare and policy analysis with nonstandard decision-makers. In P. Diamond and H. Vartianen (Eds.), *Behavioral economics and its applications* (pp. 7–77). Princeton, NJ: Princeton University Press.

———. (2009). Beyond revealed preference: Choice-theoretic foundations for behavioral welfare economics. *Quarterly Journal of Economics*, *124*(1), 51–104.

Camerer, C., Issacharoff, S., Loewenstein, G., O'Donoghue, T., and Rabin, M. (2003). Regulation for conservatives: Behavioral economics and the case for "asymmetric paternalism." *University of Pennsylvania Law Review*, *151*(3), 1211–1254.

Congdon, W. J., Kling, J. R., and Mullainathan, S. (2011). *Policy and choice: Public finance through the lens of behavioral economics.* Washington, DC: Brookings Institution Press.

Ellison, G. (2006). Bounded rationality in industrial organization. In R. Blundell, W. K. Newey, and T. Persson (Eds.), *Advances in economics and econometrics: Theory and applications* (Vol. 3, pp. 142–174). Cambridge: Cambridge University Press.

Iwry, J. M., Gale, W. G., and Orszag, P. R. (2006). *The potential effects of retirement security project proposals on private and national savings: Exploratory calculations.* Policy Brief 2006-02c Retirement Security Project.

Madrian, B. C., and Shea, D. F. (2001). The power of suggestion: Inertia in 401(k) participation and savings behavior. *Quarterly Journal of Economics*, *116*(4), 1149–1187.

Thaler, R. H., and Sunstein, C. (2003). Libertarian paternalism. *American Economic Review*, *93*(2), 175–79.

———. (2008). *Nudge: Improving decisions about health, wealth, and happiness.* New Haven, CT: Yale University Press.

CHAPTER 28

Behavioral Decision Science Applied to Health-Care Policy

DONALD A. REDELMEIER

Behavioral decision science is gaining traction and becoming a booming field. One sign of success was the award of the 2002 Nobel Prize in economics to Daniel Kahneman for work on the psychological factors that drive human decision making. Further evidence is shown by multiple best-selling books along related lines, including *Blink*, by Malcolm Gladwell and *Freakonomics*, by Stephen Dubner and Steven Levitt. Perhaps a more indirect contributor has been the contemporaneous failure of the Human Genome Project to contribute useful therapies for health care. At a root level, furthermore, behavioral decision science rests on an uncontestable medical foundation; namely, that many specific behaviors (e.g., smoking) contribute to many specific diseases (e.g., lung cancer).

Medical care is an appealing domain for the application of behavioral decision science since biotechnology is far from sufficient. Modern medical care leaves tremendous numbers of patients with ongoing suffering in nations throughout the world. Some diseases are nearly eradicated (e.g., polio), but future public-health projections are sometimes a gloomy image of the rising prevalence of chronic conditions. Even with diseases for which good treatments are available (e.g., testicular cancer), the costs of medical care can be an enormous source of societal loss. Finally, even for the few effective and simple interventions that require no specific behavior by patients (e.g., sterilized surgical equipment), medical care always requires some action by someone who judges how much effort is needed.

Clinicians generally acknowledge the role of patient behavior in health, yet almost no medical doctor has been formally schooled in behavioral decision science. As a consequence, many recommendations have been ineffective and verge on nagging patients to behave prudently. Even in recent years, the translation of behavioral insights into medical practice has been slow because the relevant material is scattered over widespread locations and because keeping up to date with biomedicine is a daunting task by itself. The intent of this chapter is to distill key concepts that appear in this book and to show the potential relevance to health policy. Where possible, specific medical applications are highlighted that seem especially plausible, feasible, or contrary to conventional practice.

Framework

Behavioral decision science is a remarkably broad field in both its specific applications and its underlying theoretical principles. This review relies on a framework based on personal clinical experience and comprising four categories: comprehension, recall, evaluation, and expression. Of course, the individual categories overlap substantially, and principles may interact at several levels. The theme is that some contributions in this volume naturally cluster together in a manner relevant to practicing physicians and that such clustering differs from the frameworks best suited to research paradigms. The ultimate contribution of behavioral decision science to medical care will depend on practical applications of selected findings to relevant patients.

Comprehension

Decision making generally requires an element of awareness, intellect, and thought. At one extreme, a patient might be declared not competent and thereby removed from the decision-making process. Such cases, however, are unusual aside from highly disabled individuals in the practice of psychiatry or neurology. Yet people's reasoning around health care is not perfect and therefore is prone to limitations in comprehension that are subtle rather than blatant. Overall, the chapters by Barr, Mullainathan, and Shafir; Pronin and Schmidt; Ross; Tyler; Ubel; and Darley and Alter

provide a broad view of how pitfalls in people's comprehension can be better understood from examining diverse fields outside of health care.

Barr, Mullainathan, and Shafir provide a case study of how credit card companies take advantage of human failings by maximizing profits at the expense of individual savings. Perhaps health insurance companies are guilty of the same unbridled profiteering. Such fundamentals translate into another argument for universal health insurance that prevails in most modern countries and against the buyer-beware health insurance that prevails for some Americans. The victimization of consumers also underscores an advantage of regulated, rather than individually negotiated, fee schedules in medical care and thereby cautions against policy proposals intended to inject more competition into physician practice and free choice by consumers. Indeed, the potential for undue profiteering in medicine is all the worse given that sick patients are scared and in no position to bargain effectively.

Pronin and Schmidt raise the idea that most people believe themselves to be unbiased, a belief that is in accord with a general theory of self-deception. Arguably, such denial of bias is a necessary and sufficient condition for errors to persist. In turn, people's blind spot toward their biases leads them to dismiss corrective procedures, discredit dissenting views, and inflame any rational discourse. The finding seems robust given its roots in people's natural tendencies toward self-enhancement, naive realism, and introspective illusions. One corrective procedure that has growing popularity in medical journals is to compel authors to declare the appearance of conflict of interest, since declaring "an appearance" is more palatable than declaring "the fact of." A more troubling implication is how introspection often hinders insight and thereby seems to endorse more peer review in bedside medical practice.

Tyler underscores the importance of cooperation in a complex world such as that required for successful medical teams. Clear examples might range from elective surgical cases, where many people need to wash their hands diligently to avoid a subsequent wound infection, to the SARS epidemic, where multiple barrier precautions need to be addressed for containing the outbreak. Throughout, the cooperation needs to be voluntary, thoughtful, and avoid crass incentives (either financial or coercive). In the future, more attention could focus on these fundamentals when attempting staff recruitment since having new members join a health-care organization requires buy-in by multiple stakeholders. Alas, a downside of cooperation is that it can degenerate into a network of cronies who perpetuate the status quo.

Ross addresses dispute resolution, a topic that has legendary importance for conflicts between clinicians and health-care administrators. Indeed, such disagreements are one of the few aspects of medical practice that are portrayed dramatically and accurately by popular television shows. Ross's summary also shows how psychology research is useful in providing a name to an everyday phenomenon and in providing a language for interpreting common sentiments. The concept of naive realism seems especially relevant to modern medicine, given the frequent arguments between MDs and MBAs. Naive realism also can explain some breakdowns that occur at the more erudite levels of medical science in dialogues between medical researchers and journal editors.

Ubel delves into the topic of why utility analysis has contributed relatively little to medical decision making despite optimistic claims in prior decades. Most reviews correctly emphasize practical barriers related to the rushed nature of clinical practice, the lack of available data on probabilities, and the fallible task of eliciting patient utilities. Ubel also correctly emphasizes some deeper ethical quandaries that would persist even if the pragmatic issues were solved. Surprisingly, clinicians do not always find the ethical discussions compelling and wonder about hidden artifacts, such as unstated assumptions about the "failure to deliver" on promises of future happiness in exchange for immediate grief. There is no argument, however, that utility analysis might still provide a helpful default for patients considering a serious choice in a medical care setting.

Darley and Alter emphasize that laws which contradict moral intuitions generate disrespect for the legal system. One parallel in medical care is thoughtless bureaucratic regulations imposed on clinicians that breed a subsequent disregard for rules and rulers. Perhaps this underpins some justification for health-care financing reform since faulty health insurers can lead to clinician subterfuge. Moreover, an affront to moral intuitions is a regular event for physicians who see patients leading prudent lifestyles (e.g., by not smoking or drinking) yet developing fatal diseases (e.g., stomach cancer). The analogue of "picking pockets at the pickpocket's hanging" also resonates with the ethical lapses by some clinical ethicists and the notorious driving habits of some experienced trauma surgeons.

Recall

Decision making generally requires background information that is separate from the immediate task yet necessary for a resolution. Few patients, for example, encounter a medical decision with no sense of information, options, or desired outcomes. Instead, each person brings to the health-care setting both

their biological profile as well as their store of beliefs, preferences, and expectations. In medicine, however, people's memories are not perfect and therefore are prone to limitations in recall. Overall, the chapters by Slovic et al.; Sunstein; Kunreuther, Meyer, and Michel-Kerjan; Fiske and Krieger; Hardin and Banaji; Steblay and Loftus; and Ellsworth and Gross provide a broad view of how pitfalls in people's recall can be better understood in diverse fields outside of health care.

Slovic et al. summarize the mismatch between the seriousness of a humanitarian crisis and the intensity of the human response, using the term *psychic numbing* to denote the general shortfall in response. Saving 80% of a small population can seem more salient than rescuing 20% of a much larger population; if so, imagine how compelling medical care becomes by saving 100% of a single life. Another related distortion that contributes to exuberant medical spending is the direct face-to-face nature of many clinical interactions, so that tears are not wiped off and the trembling hands are directly visible. All this explains heroic efforts by clinicians who sometimes work themselves to exhaustion and also illuminates the problems of the faceless uninsured throughout the United States.

Sunstein focuses on misdirected attention and misplaced emotions. The term *misfearing* is used to denote both errors, such as when people exaggerate the likelihood of an adverse event (e.g., overestimating objective risks) or magnify the importance of a consequence (e.g., increased salience). Because resources are finite, such misfearing ultimately leads to unwise choices and missed opportunities. The main issue raised by Sunstein is that cost-benefit analysis may not always be able to mitigate people's misfearing. This limitation is perhaps accentuated in medical care since rigorous data are often lacking, patients vary widely in outcomes, and the clinical arena is rushed. One caution from Sunstein is that misfearing in medicine might worsen in the future given the expanding role of the media into the previously private world of personal health care.

Kunreuther, Meyer, and Michel-Kerjan address people's tendency to underspend on protection against catastrophic risks; moreover, the compelling hundred-year historical review shows that such events are not decreasing in frequency or severity. Perhaps one corollary is to consider how communities, particularly the United States, overspend on medical care. One explanation for both patterns is that people are unmoved in the absence of tangible symptoms but can go to extremes to relieve suffering when it is salient at the individual level. Moreover, small biases become magnified because crowds become prone to imitative behavior when facing uncertain circumstances. All this provides opportunities for medical leadership since they are authorities who can change community attitudes, as demonstrated by "smoke free" hospitals and health-care centers that sponsor "fun runs" to encourage physical exercise.

Fiske and Krieger examine subtle forms of discrimination with a rigorous framework spanning the full range. In particular, the broad spectrum of discrimination can extend from hostile animus (as characterized by a belligerent bigot) to statistical differentiation (as characterized by numerical analysis of objective data). Somewhere in the middle ground of this spectrum is role stereotype (as characterized by thoughtless generalizations). This spectrum acknowledges that individuals generally respond to other people based on personal attributes and past experiences. In addition, such responses are grounded in perceptions that are categorized according to expectations (such as the presumed intentions of the other person). An awareness of this spectrum highlights how some discrimination may remain rampant in medicine because it does not follow the idealized pattern of deliberation, stability, directness, and awareness by both parties.

Hardin and Banaji discuss subtle forms of discrimination using the perspective of implicit social cognition. The core idea is that unfair institutional policies are not necessarily a reflection of bigoted leaders who are ill-motivated and misinformed. Instead, some injustice arises from implicit prejudice that extends to the target's self-identity. All this seems pertinent to medical school admission policies, particularly when identified subgroups of candidates are disinclined to accept admission offers. The review also highlights that such implicit prejudice is malleable and thereby provides some hope for future improvements. One implication, for example, is that discriminatory medical school admission patterns might lessen in future years as selection panels become more diversified.

Steblay and Loftus observe individuals wrongfully convicted (and exonerated) and trace the root cause of most (75%) miscarriages of justice to faulty eyewitness testimony. The relevance to medical care is immediate given that so much of clinical action is based on the patient's recounted symptoms and history. Practitioners, therefore, might benefit from realizing the fallibility of personal recollections and people's changing memories for actual events. Examples might include spurious accounts of chest trauma that are recalled following a diagnosis of breast cancer or assertions by patients in the aftermath of a surgical complication that they were never told of the risks. All this might help motivate alternative sources of information including computerized medical records.

Ellsworth and Gross examine psychological causes of false convictions while acknowledging that jurors are now much more aware of the possibility of making a mistake. One insight is that hindsight bias can prevail despite debiasing procedures, and, thereby,

the work underscores some cautions about medical policies aimed at assessing patient safety by analyzing adverse events. Another insight is that raising the stakes is sometimes counterproductive, in that heinous crimes can sometimes increase bias due to the increased pressure to close the case by finding someone to blame. The work also provides a helpful listing of other factors, including confirmation bias, asymmetries of information, and the absence of feedback. The corrective procedure seems fruitful, too; namely, reform the health-care system by making clinicians more willing to reconsider cases.

Evaluation

Decision making generally requires complex thought so that two or more options are compared across disparate features and an acceptable choice identified. The diversity of individual choices in medical care, for example, demonstrates how reasonable people can choose different options despite being in similar positions. In medicine, however, such divergences can be remarkably due to the underlying uncertainties, long-term consequences, and high stakes outcomes. Overall, the chapters by Wansink; Miller and Prentice; Garcia and Cohen; Jolls; and Weber provide a broad view on how pitfalls in people's evaluation of a situation can be better understood in diverse fields outside of health care.

Wansink focuses on mindless unhealthy eating, which is a topic directly relevant to medical care in endocrinology, cardiology, and several other settings. The main idea is that education is hardly the solution; instead, the main determinants seem to be size, salience, structure, stockpiling, and shapes. The fundamental point is that consumption is hard to monitor (since it disappears from view), is distorted by societal norms (that are fueled by industry), and is influenced by subtle cues (that people fail to acknowledge). The countermeasures rest on the idea that behavior modification is sometimes easier to accomplish by changing peoples' environment rather than by changing their minds. Also, the work stresses the benefit of some tailoring for each individual, such as with "Here are three foods that are best for you and three that are the worst." Ultimately, sustained changes in behavior may require accountability, such as a weighing scale for feedback.

Miller and Prentice caution that psychological interventions for changing behavior may not be as powerful as other interventions, such as passing laws (e.g., smoking bans), introducing new engineering (e.g., home diabetes monitoring), or providing economic incentives (e.g., taxing gasoline). Yet people's attitudes are malleable, so sometimes interventions can work in the long run by changing people's hearts and minds. One force is injunctive norms introduced by reviewing what most people do. In addition, clinicians should try to set straight any faulty misperceptions of the norm, such as exaggerated beliefs about college drinking. In contrast, prohibitions can sometimes lead to subterfuge and other unintended side effects. All of these approaches might be particularly helpful in promoting a healthy lifestyle among young adults and also explain the large variation in driver behavior in different countries.

Garcia and Cohen review how people sometimes underperform in educational settings due to identity threat. Perhaps this is one issue that does not directly relate to medicine since most cultures and groups have the same ideal: namely, being healthy and free of disease. Even among individuals who see themselves as "smokers," the identity is probably not a big factor in reinforcing their ongoing tobacco consumption. Perhaps patient advocacy groups might draw on such factors; for example, some "breast cancer survivors" might identify themselves with the disease and become motivated toward more community support. The underpinnings of potential corrective procedures seem valuable nevertheless, such as the need for multifactorial treatments that activate a recursive process that alters affective construal and removes critical constraining steps.

Jolls delves into employment law, given that such regulation represents one of people's most important relationships, even if it is rarely considered by physicians in medical decisions. Three nuances seem particularly relevant; namely, bounded self-interest (people will decline a salary raise if it demands chopping wood with a blunt axe), bounded willpower (people will engage in hyperbolic time discounting, to their detriment), and bounded rationality (by underappreciating risk and undervaluing disability insurance). The methodology of the Implicit Association Test seems like an intriguing way to test for latent discrimination in both work situations and medical arenas. The basic correction also seems sensible; namely, improving workplace diversity in medicine to help alleviate traditional biases by altering the surrounding environment.

Weber discusses society's hypocrisy about environmental protection and how many of the indiscretions might boil down to individual self-interest. One immediate example in health care could be hospitals, where patients and families often feel vulnerable and entitled. It is no surprise, therefore, that hospitals rarely make environmental protection a core priority. For example, few hospitals are efficient at heating and cooling, most have large staff parking lots that implicitly condone motor vehicle commuting, and all

produce vast quantities of garbage destined for landfills (including biohazards and radiation). Perhaps people's finite pool of worry nullifies deep feelings of shame. Hence, environmental crusades have generally avoided health care and gone after misdemeanors in the entertainment, leisure, and food industries.

Expression

Decision making ultimately requires action, yet the gap between reflection and action prevails in all human endeavours. Even the classic medical treatment for hypertension, for example, generally requires a sustained initiative to fill prescriptions, remember on a daily basis to take the medication, attend regularly follow-up appointments to check the adequacy of treatment, and remain open to additional considerations (such as cancer prevention or road safety). In medicine, however, the mismatch between intentions and actions results in imperfect compliance and unwanted heart attacks. Overall, the chapters by Thaler, Sunstein, and Balz; Johnson and Goldstein; Rogers, Fox, and Gerber; Fischhoff and Eggers; and Loewenstein, John, and Volpp provide a broad view of how pitfalls in people's expression can be better understood in diverse fields outside of health care.

Thaler, Sunstein, and Balz contribute a clever mnemonic for classifying choice architecture according to NUDGE factors (iNcentives, Understandable mappings, Defaults sensible, Give feedback, and Expect error). Medical care, in contrast, has focused on biology and generally neglected such fundamentals of design. The stinging indictment of U.S. Medicare Part D paperwork is a clear example of creating unwanted pitfalls and awkwardness for consumers. The optimistic interpretation is that such design lapses in medicine are accidental and would be willingly repaired by properly instructed bureaucrats. The cynical interpretation is that the administrative burdens are deliberate strategies by payors to reduce expenditures. Regardless of motivation, awareness of the principles of choice architecture can clarify what is happening in medical care financing.

Johnson and Goldstein review the large and consistent effects of defaults in guiding people's choices. Many of these effects occur because people gravitate to the path of least resistance, and such inclinations are probably frequent in medical care when patients are sick and tired. Another attractive feature of defaults is that such strategies can be exceedingly cheap, unlike educational interventions, which are costly on an ongoing basis. One emerging application of these ideas occurs in preprinted order sheets and computerized order sets that provide a convenient and legible interface for guiding physicians' treatments of patients. A further application for patient counseling might be to emphasize telling people more about base rates ("what most people in your position choose") to set a realistic initial default. The role of setting sensible defaults in medical care may likely rise in future years with the proliferation of medications and surgical techniques.

Rogers, Fox, and Gerber explore reasons why people vote and thereby juxtapose field data against the laboratory findings from experimental psychologists. The resulting attenuation of effect size is sobering and is nicely exemplified by characterizing a 6% increase in voter turnout as a "substantial" change. This calibration resonates with medical care, where the diversity and complexity of the clinical arena also makes double-digit effect sizes highly unlikely when studying changes in actual behavior. The chapter shows real savvy when recounting the multiple failures to replicate, as well as the general theme that convenience is sometimes more decisive than content. The message is that psychology has helpful insights for medical care, but nothing will cause every healthy person to donate blood or all practicing physicians to stop prescribing unnecessary antibiotics.

Fischhoff and Eggers raise the paradox that emotions both activate people through heightened motivation and hinder them by exacerbating confusion. One immediate medical analogue might be to affirm the traditional dyad of the suffering patient (who is therefore highly motivated) and the tranquil physician (who can therefore think straight). Hence, the doctor-patient relationship is a powerful team that will not be obsolete any time soon. Another affirmation is toward public health agencies in channeling community emotions into constructive work, as was positively demonstrated during the SARS epidemic and less positively in campaigns toward prostate cancer awareness. The caveat is that the gap in science is filled in by the imagination, which creates distinct stumbling blocks when people view themselves as experts (e.g., on nutrition, exercise, driving).

Loewenstein, John, and Volpp focus on some potential mitigating strategies and, in particular, counterbiases for correcting people's biases. All of this is relevant to health care, such as the classic example that people's loss aversion is partially offset by their natural tendency toward projective optimism. Indeed, skilled clinicians often have an intuitive sense of such strategies when eliciting informed consent or guiding shared decision-making choices. The work of Loewenstein, John, and Volpp highlights the specific potential for providing financial incentives to patients as a way to improve compliance with medications; however, the high costs of current medical care may

make payors unwilling to fund such prospects unless net cost savings are quickly apparent. Loewenstein et al. also suggest greater feedback directly to patients about lab-test results and scheduling; on balance, these innovations around information systems seem relatively more attractive by reducing the bottlenecks that occur in doctors' offices.

Summary

Behavioral decision science is a broad field, and this summary attempts to distill some major points contained in this book that are relevant to medical decision making. The largest gap in knowledge is to determine which concepts are "large" or "small" in the medical arena. The laboratory studies in psychology are, of course, insightful, but the results do not immediately translate to patient care because the findings are based on volunteer subjects facing hypothetical tasks with somewhat artificial outcomes. The field studies in psychology and related fields seem to show more modest results when they are explored in a more complex setting. The priority, therefore, should be to popularize this body of knowledge so that selective strategies can be explored for future medical policy making.

CHAPTER 29

Quis custodiet ipsos custodes?

Debiasing the Policy Makers Themselves

PAUL BREST

Knowledge of the judgment and decision-making (JDM) biases discussed in this book can be applied to the behavior of citizens, consumers, organizations, and policy makers. Many of the essays inform policy makers about how to use this knowledge either to mitigate the biases of individuals and organizations (e.g., to prevent discriminatory behavior) or to manipulate inevitable biases so that people act in their own or society's best interests (e.g., make appropriate investments for their own futures or protect the environment).

My discussion will focus specifically on the behavior of the policy makers themselves, with the aim of mitigating biases and other errors in their own decision making. I will not address the perfectly legitimate concern that policy makers might manipulate individuals' biases for self-serving or corrupt ends; rather, my target is the policy maker who seeks to act in the public interest but whose judgments may nonetheless be biased—typically without self-awareness (Pronin and Schmidt, this volume). Although much of what follows will focus on *biases* that typically arise from unconscious processes, it also touches on some plain judgmental *errors*, for example, errors in using statistics. (For a good review of problems of both sorts, see Larrick, 2004).

What Constitutes Nonbiased Behavior?

The concept of a bias assumes some ideal of unbiased judgment and decision making. As the two components of JDM suggest, one can be biased in making empirical judgments (predictions as well as findings of past or present facts) and in making decisions, or choices. Without overly complicating matters, the ideal against which judgmental biases are compared is seeing the world as it actually is, even when judgments must be probabilistic and based on "multiple fallible indicators" of reality (Hammond, 2000).

The unbiased ideal for making decisions or choices is somewhat more complicated. Before turning to policy makers, it is helpful to consider a model relevant to individuals acting in their private capacity and to look briefly at the role of counselors in debiasing individuals.

Individuals

Much of the literature on JDM and behavioral economics invokes a model that equates nonbiased decision making with the maximization of subjective expected utility (SEU). The modifier *subjective* captures the understanding that preferences differ from one person to another—in realms ranging from food to religion to tolerance of ambiguity and risk—and that SEU does not make value judgments about these preferences.

Biases are the cognitive and motivational phenomena that lead individuals to systematically make suboptimal decisions in terms of their experienced utility. Although using SEU as the desideratum for individual decision making has its share of conceptual as well as empirical problems (Keys and Schwartz, 2007), I do not know of any successful efforts to replace it with an alternative model.

Counselors

Individuals often seek the advice of lawyers, physicians, stock brokers, and other counselors. Subject to legal and ethical constraints, the essential function of counselors is to promote the SEU of the individuals who engage them.[1]

What should counselors do when they believe that clients' intended decisions are the result of biases and therefore compromise their SEU? Korobkin and Guthrie (1997) make a good case for a "cognitive error approach to counseling," which calls for

respecting their clients' preference structures but nevertheless still helping them recognize and counter cognitive biases. Implicit in Korobkin and Guthrie's view is that counselors have a perspective not available to their clients (cf. Pronin and Schmidt, this volume). This may be true either because counselors possess expertise in decision making or because they are disinterested.

Policy Makers

Policy makers include legislators, administrative officials who determine facts or make decisions, and judges (notwithstanding the unconvincing claims of federal judicial nominees that they do not make policy). Within the bounds of their authority, policy makers serve "the public interest." Whatever this capacious term means, it reflects the fact that, unlike individuals acting for themselves, policy makers are concerned not with their own, but instead with others' utilities. They also differ from counselors acting on behalf of clients: policy makers are responsible to multiple stakeholders, typically with divergent, and often competing, interests.

Policy making requires accommodating or choosing among the heterogeneous interests of individuals whose utilities differ in tastes, risk attitudes, and fundamental values. Policy makers cannot make decisions by aggregating or averaging their constituents' interests but instead must make distributional judgments that promote some people's welfare at the expense of others.[2] Indeed, a classic problem arises from the tradeoff between the goals of maximizing aggregate welfare and reducing inequality (Rawls, 1971).

One might argue that the policy maker's task of representing multiple constituents is just an extension of an individual decision maker's task of reconciling his or her own multiple interests. After all, an individual must deal with conflicts among her own interests and values, between her present and future selves, and even (say, when engaging in estate planning) with lives not yet in being. But whatever the conceptual similarities may be, the practical tasks of determining the SEU for a polity are immensely more complex.

Although SEU as such does not provide a plausible reference point for assessing policy makers' biases, cost-benefit analysis offers a procedure that identifies who will benefit how much from a regulation (e.g., standards for workplace safety) and who will pay the costs, taking account of risk and discount rates with respect to future benefits (Sunstein, this volume).

In any event, the same deviations from rational decision making that are likely to reduce an individual's experienced utility will signal suboptimal policy making. The major questions that a counselor would ask to assess whether an individual client was undermining his or her own SEU apply to policy makers as well (Brest and Krieger, 2010, p. 386):

1. Is the decision based on incorrect data or the incorrect analysis of data?
2. Is the decision based on inadequate consideration of the interests at stake?
3. Does the decision violate one of the axioms of expected utility—for example, transitivity or procedural invariance?[3]
4. Is the decision sensitive to the way the issue is framed and made in a context where the framing is highly variable or manipulable?
5. Is the decision sensitive to affect and made in a context where affect is highly variable or manipulable or influenced by factors unrelated to the decision?
6. Is the decision maker subject to undue social influence?

The remainder of this essay will use such deviations from *procedural rationality* as the reference point for biased decision making of any sort.[4]

A Spectrum of Policy-Making Procedures

In the following section, I will discuss the biases that attend three policy-making functions: adjudicative fact-finding (concerning facts about a particular event), legislative fact-finding (concerning facts that underlie policies), and choice, or decision making. All three of these functions may be performed through more or less formal procedures, with some formal procedures reducing the opportunities for judgmental and decision-making biases.

The judicial trial is paradigmatic of procedures at the formal end of the spectrum. In determining facts, an independent judge hears testimony from all parties to a dispute. The proceedings are subject to rules of evidence designed to exclude testimony that lacks probative value (and even to exclude probative evidence that is likely to create unfair prejudice), and testimony is subject to cross-examination. Judges typically explain their factual and legal determinations in written opinions. Their decisions are subject to review by appellate courts, which comprise a number of members who write majority, concurring, and dissenting opinions. In a word, the judicial process is characterized by argument.

The procedures of many administrative agencies mirror those of the courts, whereas others are far less formal. Commissions, such as those inquiring into the events of 9/11 or the BP Gulf of Mexico oil spill,

typically use an inquisitorial process, in which members of the commission interrogate witnesses, rather than the adversarial process characteristic of American courts. Like courts, commissions justify their conclusions in writing.

Legislatures lie at the informal end of the spectrum. Individual legislators—whether as members of committees or of the legislature as a whole—engage in implicit legislative fact-finding and explicit decision making. But they are not bound by rules of evidence nor even by a requirement of stating, let alone justifying, their conclusions. Representative Jones may vote for a climate bill, and Representative Smith may vote against it, based on their different implicit beliefs about the underlying science, and even for reasons that they might prefer not to articulate publicly.

These differences in procedures have implications for the policy makers' vulnerability to biases and other errors. In particular, relatively formal procedures have these characteristics:

Rules of evidence highlight some potential biases, at least bringing them to the attention of the fact finder and sometimes averting them.

The presence of advocates for the parties and other stakeholders increases the likelihood that decision makers will consider all of the relevant facts and interests. (The presence of advocates also helps counter some biases by giving opposing parties the incentives to produce alternative scenarios or reference points by which gains and losses are framed.)

The practice of justifying decisions in writing subjects intuitions and prejudices to a degree of analytic scrutiny by the policy maker himself as well as by others. (In the year that I clerked for Supreme Court Justice John M. Harlan, he changed his vote in several cases after attempting to write an opinion and concluding "It just doesn't write convincingly.")

Multimember decision-making bodies typically aim to achieve consensus, which requires considering the views of colleagues and attempting to persuade them to one's position. This goal has the potential to produce less-biased outcomes but also makes decisions vulnerable to the social influences of peers (Sunstein et al., 2006).

Federal judges and some state judges are not subject to removal. For better or worse—but mostly for the better—this reduces the likelihood that their decisions will be based on their predictions of how they will be held accountable (Lerner and Tetlock, 1999; Siegel-Jacobs and Yates, 1996).

To What Biases Are Policy Makers Susceptible, and Can They Be Debiased?

Everyone interested in a particular policy outcome—whether the lawyers representing parties to litigation, paid lobbyists, or civil society organizations—also has an interest in influencing the policy makers responsible for the outcome, and many advocates are skilled at doing so. Policy makers have an interest in avoiding being (or appearing to be) biased, whether through the influence of others or from the biases they bring from their own perceptions and experiences. The difficulty lies in the "bias blind spot" (Pronin and Schmidt, this volume)—the fact that we are unaware of our biases most of the time. (Generally, if we are aware of them and continue to act on them, we do not think of them as biases.)

Before turning to specific biases and strategies for mitigating them, it is useful to review the dual-process, or two-systems, model of cognition, referred to in many of the chapters of this book, and to address two different strategies for dealing with biases. System 1 is intuitive, unconscious, automatic, and fast. System 2 is analytic and conscious; it is cognitively effortful and works more slowly (Kahneman and Frederick, 2002). While System 1 plays an essential role in judgment and decision making, its reliance on schematic processing and heuristics gives rise to errors—the stock and trade of the JDM research agenda.

There are two essentially different strategies for addressing the biases that flow from System 1. *Debiasing* involves the relentless application of System 2 rationality. Cass Sunstein's proposal for cost-benefit analysis is paradigmatic (Sunstein, this volume). *Counterbiasing* counters one System 1 phenomenon with another. The behavioral-economics-oriented essays of Barr, Mullainathan and Shafir; Thaler, Sunstein and Balz; and Loewenstein, John, and Volpp (all in this volume) are paradigmatic of this strategy. Their task is to show how "a range of decision phenomena that are typically viewed as errors—including the default bias, loss aversion, present-biased preferences, and nonlinear probability weighting—can be exploited to devise interventions to help people accomplish their goals" (Loewenstein and Volpp, this volume). (For a discussion of both debiasing strategies, see Milkman, Chugh, and Bazerman, 2009.)

The distinction between debiasing and counterbiasing is not always sharp. For example, while countering the abstractness and uncertainty of the future harms of global warming by instilling fear is a System

1 strategy, making future harms more concrete can also improve the cognitive processes of System 2 (see Weber, this volume).

Debiasing through rational analysis is transparent and it invites argument and discussion. But fighting one System 1 fire with another may sometimes be the only effective method of countering biases and can also be transparent to the decision maker. For example, legislators might consciously and intentionally follow Hardin and Banaji's strategy of seeking contact with people of different races, ethnicities, or classes to mitigate their own negative stereotypes and empathize with those adversely affected by proposed policies (Hardin and Banaji, this volume).

I now turn to the particular biases to which policy makers are susceptible and to the possibility of preventing or mitigating them. While the activities of adjudicative fact-finding, legislative fact-finding, and decision making sometimes overlap and are vulnerable to some common biases, they have different centers of gravity, and it is therefore useful to consider them separately.

Adjudicative Fact-Finding

Adjudicative fact-finding seeks to determine past events and sometimes to assign responsibility for them. The paradigmatic example of adjudicative fact-finding is a trial judge's or a jury's determination of whether the accused killed the victim. Not just courts but state, local, and federal administrative agencies engage in adjudicative fact-finding—for example, determining whether a particular land use violates a zoning ordinance or whether a manufacturing plant is emitting excessive pollutants. Intelligence analysts also engage in adjudicative fact-finding, piecing together various data to determine, say, changes in political power in North Korea or whether Iraq has weapons of mass destruction.

PERCIPIENT WITNESSES

Contested issues of fact often depend on the testimony of percipient witnesses—people who observed the relevant events. These witnesses' perceptions, or their memories of what they perceived, may be biased in any number of ways, including:

Errors in perception. The possibilities of errors begin with witnesses' perceptions of an event (Ellsworth and Gross, this volume). Simple perceptual mistakes about who or what we saw are influenced by schematic expectations, which in turn can be influenced by partisanship (say, for a particular football team, Hastorf and Cantril, 1954) or by unconscious stereotypes about people of different races and genders (Hardin and Banaji, this volume; Fiske and Krieger, this volume).

Distortions of memory and retrieval. Many internal thought processes and external stimuli intervene between the moment of a witness's perception and her testimony in a formal proceeding (Loftus, 1996; Schachter, 2001). Suggestive questioning by police and poorly designed lineups are common examples of error-inducing police procedures (Steblay and Loftus, this volume).

Considerable effort has gone into designing interrogation and identification procedures that reduce errors (Steblay and Loftus, this volume). Mitigating the implicit prejudice that infects schematic processing is a much more difficult task. It requires changing people's attitudes, which, to put it most optimistically, is a long-run strategy.

ADJUDICATORY PROCEDURES

Fair adjudicatory procedures provide an opportunity to expose and counteract some of the biases of percipient witnesses. Cross-examination is a key way to test a witness's perception and memory—granted that the adversarial "preparation" of witnesses by their lawyers can produce convincing but untrue testimony (Ellsworth and Gross, this volume).

Expert testimony can point up the susceptibility of particular identification procedures to bias. In addition, the holding of *Daubert v. Merrell Dow Pharmaceuticals* (1993), prevents decisions based on "junk science." In that case, which involved the claim that the drug Bendectin was responsible for birth defects, the Supreme Court held that federal courts could not rule for the plaintiff based on an expert's testimony unless the underlying research was accepted by the relevant scientific community.

But adjudication has its own biases.

Hindsight Bias

Hindsight bias is a common problem when a judge or jury must determine, after an injury or other liability-inducing event has occurred, whether a party took reasonable precautions to prevent its occurrence. Although this phenomenon is notoriously difficult to debias once a fact finder actually knows the outcome (Fischhoff, 1975), some judicial rules tend to protect against it—for example, the plaintiff in a slip-and-fall case cannot introduce evidence that the defendant subsequently repaired the property to make it safer (Kamin and Rachlinski, 1995).

Anchoring and Insufficient Adjustment

Especially where numbers are involved, as in civil claims for pain and suffering, both parties may influence the fact finder by providing anchors for the award of damages (Guthrie, Rachlinski, and Wistrich, 2001; Korobkin and Guthrie, 1994; Malouff and Schutte, 1989). Even extravagant anchors may have an effect that is not easily debiased.

Confirmation Bias

The process of fact-finding often starts with a hypothesis about the conclusion, after which the fact finder may systematically favor evidence that supports the hypothesis (Ellsworth and Gross, this volume). The phenomenon is hardly limited to judicial disputes. Confirmation bias may have contributed to the intelligence community's incorrect conclusion that Saddam Hussein possessed weapons of mass destruction. Perhaps the strongest debiasing tactic is a requirement that one justify one's conclusion orally or (better yet) in writing, describing the evidence on both sides. While some adjudicatory procedures require this, others, including jury verdicts, do not—although jurors tend to justify their positions to each other in the jury room.

Treating Evidence Originating from a Single Source as if It Were Based on Multiple Independent Sources

Fact finders may treat multiple pieces of evidence as confirming a conclusion without examining whether they originated from a common source. Intelligence analysis is particularly vulnerable to this error, which again may have played a role in the conclusion that Iraq had weapons of mass destruction. Availability cascades (Sunstein and Kuran, 1999) also induce this error: one person's claim that, say, residents of an area are suffering illnesses because of toxic wastes gets repeated multiple times, not only by people who believe it but also by some who may have their doubts. Formal evidentiary rules tend to prevent this sort of error in courts (e.g., restraining public officials who may want to appear responsive to constituents). Other fact finders must develop their own ways to guard against it.

Difficulties in Prediction

Even experts are not very good at predicting the consequences of interventions in complex systems because those systems often behave in nonlinear and chaotic ways. If the "butterfly effect" suggests that the perturbations of a butterfly flapping its wings in Indonesia may lead to a hurricane in Florida, I will coin the phrase "elephant effect" to describe the unpredictable effects of, say, the invasion of Afghanistan or the adoption of massive regulatory schemes. Edmund Burke said as much in *Reflections on the Revolution in France* (1790), and Phillip Tetlock essentially confirmed this view in *Expert Political Judgment* (2005). Whether or not this counsels Burkean conservatism—after all, doing nothing is also a decision—it does call for doing one's best to anticipate the unintended consequences of a decision. "Adversarial" decision processes, such as using "red" and "blue" teams, can aid this process. Group dynamics that press for too-quick a consensus do just the opposite.

On a much more local scale, policy makers are famous for committing the planning fallacy. Flyvbjerg, Bruzelius, and Rothengatter (2003) described the pervasive cost overruns and delays of large-scale transportation and other construction projects. Although the causes lie mostly in System 1, solutions may be found in System 2 processes, such as carefully thought through GANTT charts, fault trees, and "premortem" analyses (Klein, 2007) that try to identify everything that can go wrong at each stage.

RELIANCE ON CLINICAL RATHER THAN ON STATISTICAL PREDICTION

Officials are often called upon to predict individuals' behavior—for example, a parole board is required to predict the likelihood of recidivism of a prisoner petitioning for early release. Numerous studies suggest that in situations where prediction by any means tends to be inaccurate, decision makers over-rely on their intuitions rather than on, say, simple linear regression models that are more accurate (Dawes, Faust, and Meehl, 1982).

In any event, accurate prediction depends on getting clear and timely feedback about the accuracy of one's previous predictions. Because of the focused and repetitive nature of their task and the availability of clear and timely feedback, weather forecasters make quite accurate predictions. By contrast, many policy makers are faced with decisions of a nonrepetitive, if not unique, nature and have poor feedback, so their predictions are quite poor (Tetlock, 2005). This may just be the nature of the beast.

Legislative Fact-Finding

While adjudicative fact-finding typically focuses on particular events, legislative fact-finding determines facts about the physical or social world that underlie regulations and other legislative actions. Legislative facts typically are used to predict whether a proposed policy will work as intended. For example, policies addressing climate change are—or should be—informed by factual determinations about the relationship between greenhouse gas emissions and global warming, the effects of global warming on precipitation and sea

levels, and the environmental and economic consequences of particular regulatory schemes. Legislative fact-finding underlies many policies, ranging from health and safety regulations to consumer protection legislation. Appellate courts also engage in legislative fact-finding—for example, in concluding that segregation inflicts psychic harms on African Americans (*Brown v. Board of Education*, 1954) or that the death penalty is not applied in a racially discriminatory manner (*McCleskey v. Kemp*, 1987).

Legislative and adjudicative fact finding sometimes overlap. For example, the issue in the *Daubert* case (mentioned above) could have arisen in the context of the Food and Drug Administration's decision to regulate a class of pharmaceuticals rather than in a tort case brought by an injured party. Legislative fact-finding is subject to many of the same errors, as well as others.

POOR GRASP OF PROBABILITY, STATISTICS, AND EMPIRICAL METHODOLOGY

Much of the evidence underlying legislative facts is probabilistic in nature. And much research in medicine and the natural and social sciences relies on experiments or econometric analyses, where an outcome can only be described in terms of statistical significance, and where the validity of conclusions depends on complex and sometimes contested questions of methodology. Even when judges hear expert testimony on these matters, they often find the conflicting views difficult to untangle—and many policy makers don't have the advantage of experts.

In a famous case, a California court convicted a man and a woman for robbery based on the confusion between the conditional probabilities P(A|B) with P(B|A) in determining the likelihood of a couple in Los Angles having their particular personal characteristics (*People v. Collins*, 1968). NASA's disastrous decision to launch the Challenger space shuttle was partly rooted in errors of data presentation and statistics in which the engineers considered only the temperatures at which the O-rings failed but not the temperatures at which they did *not* fail (Vaughan, 1997).

Training in statistics reduces some judgmental errors (Lehman, Lempert, and Nisbett, 1988), and there have been efforts to train judges in statistics (Federal Judicial Center, 2000). But policy makers—from judges to administrative officials to legislators—usually have little education in these matters. Indeed, some may share Mark Twain's view that there are "lies, damned lies, and statistics" (Twain, 1906–1907).

AVAILABILITY AND RELATED BIASES

Ignorance of probability doubtless contributes to the availability bias and the broader phenomenon of "misfearing," described by Sunstein's essay in this volume. But even the statistically educated tend to have distorted views of risk in the face of vivid events, such as a terrorist bombing or shark attack. Individual policy makers' own vulnerability to the availability heuristic is inevitably reinforced by their constituents' perceptions. (Many more people are killed by falling television sets than by shark attacks. As a journalist noted, tongue-in-cheek, " 'watching "Jaws" on TV is more dangerous than swimming in the Pacific' " (*New York Times*, 2001). But it is shark attacks that make the news and to which policy makers feel they need to respond.)

People are prone to other risk-related errors as well. They have difficulty understanding the meaning or policy implications of very low probabilities and different ways of representing probabilities—e.g., 1 out of 1,000,000 times versus .000001—can produce different emotional responses to the same situation (Blumenthal, 2007; Kunreuther, Novemsky, and Kahneman, 2001). More generally, people are subject to probability neglect, entirely ignoring low-probability risks, being insensitive to a broad range of differences in risk, and being readily subject to alarm and hence to overestimating the danger of moderate risks. Responses to risk are also affected by factors not necessarily related to rational decision making—for example, whether the risk occurs in nature or is anthropogenic. With respect to debiasing, Sunstein (this volume) persuasively argues that the essentially System 2 procedure of cost-benefit analysis presses fact finders to consider actual risks with far more precision than reliance on System 1 intuitions.

AFFECT HEURISTIC

Our judgments of risk are often based more on intuition than on dispassionate analysis (Slovic et al., 2002), and we tend to view activities deemed beneficial as less risky than those that are not (Alhakami and Slovic, 1994). While the obvious debiasing technique is cost-benefit analysis, its treatment of risk has been criticized as insufficiently responsive to the different cultural and world views that give rise to the intuitions (Kahan et al., 2006).

PSYCHIC NUMBING

Slovic and his colleagues write (this volume) that even though our moral theories hold that every life is equally valuable, our moral intuitions fail to respond to the scale of massive deaths, whether by mass atrocities or natural causes. This phenomenon of psychic numbing, or the collapse of compassion, is captured by the comment, attributed to Joseph Stalin, that "one death is a tragedy; a million is a statistic" and

is manifested in the inadequacy of institutional responses, say, to genocide. Ironically the very documentation of the statistics, which is part and parcel of System 2 analysis, undermines the compassionate intuitions of System 1. Slovic et al. suggest approaches for debiasing that activate the affective imagery to which System 1 responds or that promote relentless System 2 deliberation.

OVERCONFIDENCE, MOTIVATED SKEPTICISM, AND CONFIRMATION BIAS (REDUX)

Policy makers may assign a smaller-than-warranted confidence interval to their factual conclusions. Overconfidence in adjudicative fact-finding is mitigated by procedures that require judges, juries, and administrative agencies to hear both sides of an argument, which implicitly incorporate the most effective debiasing of overconfidence: "Consider why your estimate may be wrong." (Burdens of proof, such as "beyond a reasonable doubt" or "clear and convincing evidence," provide asymmetrical limits on overconfidence.)

In legislative fact-finding, overconfidence combines with motivated skepticism, confirmation bias, and the gravitational force of prior commitments to make it particularly difficult for policy makers to be open to considering alternative positions relevant to major policy issues ranging from climate change to the right to carry concealed weapons.

Motivated skepticism refers to people's tendency to be less critical of facts and arguments that support their preferred result than a dispreferred one (Ditto and Lopez, 1992).

I mentioned confirmation bias in the section on adjudicative facts. Legislative fact-finding is peculiarly vulnerable to the distortions of naive realism and the phenomenon of biased assimilation—including questioning the motives of those having different views (Ross, this volume). A chastening experiment, in which participants' prior views about the deterrent effect of the death penalty were reinforced by hearing conflicting evidence, provides a vivid example and also suggests the limits of the consider-the-opposite approach to debiasing (Lord, Ross, and Lepper, 1979).

Even at its best, legislative "debate" is just that, with legislators (understandably) not paying any attention to one another's prepared speeches. Any real discourse takes place outside of the chamber, but there are few occasions for informal bipartisan discussion. Lee Ross's chapter in this book, which focuses mainly on the role of mediators in international disputes, offers some possibilities for improving domestic policy making as well: given the opportunity, mediators can help partisans engage in what he describes as the "obvious antidote to naive realism and its attributional consequences—that is, the open, sustained, sympathetic sharing of views and perspectives" through "dialogue in which they talk about their factual assumptions and the complexity of their values, rather than simply defending their positions." The Aspen Institute's Congressional Program and the Wilson Center on the Hill (Woodrow Wilson International Center for Scholars), which educate members of Congress on substantive matters, and James Fishkin's deliberative polls are suggestive of the possibilities—although at this particular time, the barriers seem greater than ever (Fishkin, 2009).

BIASES IN SOCIAL PERCEPTION: STEREOTYPING

Policy makers' social stereotypes generally reflect those of other members of society. This seems most evident in the enactment of, or failure to repeal, laws that discriminate against people because of race, ethnicity, gender, sexual orientation, and other personal characteristics. Social stereotyping may also be manifest in laws that, although neutral on their face, have a disproportionate impact on negatively stereotyped groups. For example, the heavier sentences imposed for crimes related to crack than to powder cocaine may reflect legislators' association of crack with a lower class Black lifestyle and powder cocaine with a less threatening White lifestyle.

Debiasing requires being motivated to address unconscious biases, actually recognizing them, and making efforts to counter them (Wilson, Centerbar, and Brekke, 2002)—quite significant obstacles.

DISTORTIONS OF FACTS BY ADVOCATES

Advocacy in adjudicatory bodies is highly constrained by formal procedures that tend to mitigate the parties' efforts to distort the facts in their favor. Interest groups often try to influence legislators through direct lobbying or by appealing to constituents through the media, and they use every technique known to marketing to make their messages stick. (In contrast to vivid images and sound bites, statistics are pallid and have no emotional power. It is notoriously difficult to counter these biases, in part because fact finders are unaware of their effect.)

LEAKAGE OF TRADE-OFF PREFERENCES INTO FACT FINDING

In principle, fact-finding should precede making choices and trade-offs. In practice, the anticipation of trade-offs and aversion to certain outcomes may infect the fact-finding process. A legislator who believes that the regulation of greenhouse gas emissions will cause unemployment is likely to undervalue evidence

of the harms of global warming. A robust debiasing procedure would separate the fact finder from the decision maker, something that does not often occur with legislative decisions.

Decision Making

With legislative facts in hand and an appreciation of the uncertainties, policy makers must consider how best to address a problem; for example, if the goal is to reduce carbon dioxide emissions, should they tax carbon emissions, subsidize solar, wind, or nuclear power generation, or use a combination of both taxes and subsidies? This is the paradigmatic task of legislatures. It requires considering how alternative solutions affect different stakeholders and also requires making trade-offs among stakeholders' interests. The process is subject to a variety of biases and other barriers to sound decision making.

DEFINING THE PROBLEM AND CONSIDERING SOLUTIONS

Several common decision-making defects result from a combination of bounded cognition and imagination: failing to consider all the important interests or values, defining a problem too narrowly, and homing in on a single, attractive solution without considering others that may better satisfy stakeholders' interests. The main barriers are inattention, impatience, time pressures, and environments that stifle creativity. These barriers can be overcome by creating processes that systematically canvass the relevant interests, values, solutions, and cultures that nourish creativity.

REASON-BASED AND VALUE-BASED DECISION MAKING

As its name suggests, a reason-based process involves giving reasons for or against proposed outcomes. This is the core of argument, whether in courts and administrative agencies, academia, or personal life. However, as Shafir demonstrated, reasoned-based decision strategies are vulnerable to framing: in a classic experiment involving a custody dispute, participants made different choices depending on whether their task was to "award" or "deny" custody; the way the issue was framed caused the subjects to focus respectively on the parents' positive or negative attributes (Shafir, 1993).

Value-based decision processes identify the interests at stake, assign (sometimes weighted) values to them, and favor the outcome that optimizes the various values. There are numerous formal methods of value-based decision making, most of them under the rubric of multi-attribute utility theory (MAUT). Administrative policy makers sometime use value-based decision procedures, for example, in determining the location of a power plant or routes for transporting nuclear waste (Keeney, 1996). These methods have the advantage of systematically canvassing all the interests and are highly transparent. But not all decisions can be made solely through this process.

SEQUENTIAL CONSIDERATION OF CHOICES

A value-based process ensures consideration of a number of plausible alternatives and presses the decision maker to make trade-offs among them. Many actual policy decisions result from the consideration of a highly constrained set of options that may be voted up or down without comparing them to alternatives. Much legislation—especially in times of great political polarization—has this generally suboptimal character (Milkman, Chugh, and Bazerman, 2009).

CHOICE OVERLOAD

When faced with too many choices, individuals make suboptimal decisions, eschewing systematic consideration of the alternatives (Iyengar and Lepper, 2000). There is no reason to think that policy makers are immune to this phenomenon. In contrast to individuals making one-off decisions, however, policy makers involved in recurring choices of the same sort are able to develop systematic choice processes, for example, along the line of MAUT.

CONTEXT DEPENDENCE

Like consumers, policy makers choosing an option may be influenced by the presence of options that are more or less attractive (Kelman, Rottenstreich, and Tversky, 1996). Consider, for example, extremeness aversion: an official presented with the choice of purchasing one of three differently priced properties for a municipal park will tend to opt for the middle-priced one, especially because, if the choice turns out badly, it seems easier to explain to constituents that one chose the "reasonable" middle ground (Guthrie, 2003a).

THE EFFECT OF THE DECISION MAKER'S EMOTIONAL STATE

Decision outcomes may be affected by the decision maker's emotions at the time of making a decision, whether or not the emotions are relevant to the outcome (Loewenstein and Lerner, 2003). Anger or elation arising from one's personal life can affect a major domestic or international policy decision (Forgas, 1995). Lerner and her colleagues differentiated between such "incidental" emotions and "integral" emotions, which arise from the issue at hand and can provide valuable information about the merits of the decision (Barlow, 1998, "anxiety is the shadow of intelligence"; Zimmerman and Lerner, 2010). Lerner

et al. proposed self-debiasing strategies to reduce the effects of incidental emotion: diagnosing one's emotions, absorbing others' perspectives, and understanding the uniqueness of the situation. And they also suggest that accountability for the decision process can provide an external check (Lerner and Shonk, 2010).

AVOIDING DIFFICULT TRADE-OFFS

Policy makers do not like making difficult trade-offs, especially when they are held accountable for outcomes. When pressed, they engage in "buck-passing, procrastination, and obfuscation" (Tetlock, 2000, p. 240). And when responsibility for a decision cannot be avoided, they tend to "play down the strengths of the to-be slighted value and play up the strengths of the to-be selected value" (Tetlock, 2000, p. 245; Tetlock and Boettger, 1994). This is a difficult phenomenon to counteract other than by replacing accountability for outcomes with accountability for following a sound process (Lerner and Tetlock, 1999; Siegel-Jacobs and Yates, 1996)—not an easy sell to most political constituents.

LOSS AVERSION, BEHAVIOR UNDER RISK, AND ACTION AND INACTION BIASES

Individuals are loss-averse; they tend to value an entitlement they currently possess more than one that they don't have. And they are risk averse when a decision is framed in terms of gains, and risk seeking when it is framed in terms of losses. Policy makers doubtless also exhibit these attitudes. After all, the famous Kahneman and Tversky epidemic hypothetical is a policy problem (Tversky and Kahneman, 1981).[5]

Almost every social policy intervention triggers the Burkean uncertainties of unintended consequences. Consistent with the general tendency to regret actions more than inactions (Miller and Taylor, 1995), policy makers are inclined to adhere to the status quo—although in public crises, they may exhibit an action bias in response to the felt pressure to "not just stand there but do *something*." (The enactment of the Sarbanes-Oxley Act in response to the scandals at Enron and other corporations may be an example.) And policy makers are prone to take undue risks when falling short of a target to which they are held accountable (Guthrie, 2003b; Rachlinski, 1996). NASA's decision to launch the Challenger Shuttle at a temperature known to make the O-rings fail provides a tragic example (Vaughan, 1997).

These phenomena (largely described by prospect theory) are pervasive but are amenable to counterbiasing. For example, in addressing the tendency to underinsure against catastrophic harms, Weber argues that "by focusing . . . attention on the severity of the possible loss and resulting consequences, all smaller losses (including the insurance premium) are to the right of this new reference point, making this a decision in the domain of (forgone) gains, where people are known to be risk-averse and will choose the sure option of buying the insurance" (Weber, this volume). The biases can also be mitigated by cost-benefit analysis, which replaces subjective perceptions of gains, losses, and risk with cold numbers (Sunstein, this volume).

COGNITIVE MYOPIA AND THE EXCESSIVE DISCOUNTING OF FUTURE BENEFITS

When considering social or environmental investments, such as whether to incur costs today to improve education, health, or the environment tomorrow, the immediate costs seem far more concrete than future benefits or the future costs of inaction. Moreover, policy makers as well as individuals often apply extravagant discount rates to future benefits and exhibit a strong present bias (Ainslie and Haslam, 1992; Giddens, 2008). Accountability to present constituents, who likely exhibit the same biases, only exacerbates the tendency. Counterbiasing might involve making the future more vivid (System 1) or asking policy makers to compare their implicit discount rates with the interest they could get from a bank (System 2).

A related question arises in the allocation of public resources between aiding the current victims of an epidemic (AIDS) or disaster (flooding) versus preventing or mitigating future harms. As Kunreuther notes, legislatures are likely to spend extraordinary resources on the former and shortchange the latter (Kunreuther et al., this volume) While political pressures play a major role, legislators and constituents alike are afflicted with cognitive myopia, which is exacerbated by its close cousin, the identifiable victim effect (Jenni and Loewenstein, 1997).

Cost-benefit analysis is a potentially effective debiasing strategy because it requires consideration of the costs, benefits, and risks at a granular level rather than in a gestalt framework that invites System 1 biases. Indeed, Sunstein argues that "if the public demand for regulation is likely to be distorted by unjustified fear, a major role should be given to more insulated officials who are in a better position to judge whether risks are real" (Sunstein, 2005, p. 126; but see Kahan et al., 2006).

Social Influences

The preceding discussion has focused on cognitive errors and biases. But policy makers are also subject to biases stemming from social influence.

SOCIAL PROOF, OR CONFORMITY

The main way that someone new to an organization learns its norms is by observing the behavior of colleagues. Someone who joins a public agency that has high ethical standards is more likely to act honestly than her counterpart who joins an agency where corruption is widespread. Some legislative bodies have traditions of pork-laden earmarks, whereas this is not acceptable practice in others. The best insulation against being drawn into dubious practices is a strong inner moral compass, supplemented by trusted advisors or counselors outside the organization.

RECIPROCITY

We tend to accede to requests from people who have done favors for us—even when the favors were uninvited and trivial and the subsequent requests are substantial (Cialdini, 2008). Some government agencies address this problem through prophylactic rules that forbid employees from accepting even a free cup of coffee. At the other extreme are the log-rolling, horse-trading, and close relationships with lobbyists common in legislatures. It might be that these practices are so much the stock-in-trade of legislators that they are less vulnerable to unconscious influence than the rest of us. (I doubt it.)

ESCALATION OF COMMITMENT

Once we have taken a position, we tend to act in a manner consistent with it, even when the subsequent actions are not in our own or society's interests. This consistency postpones, even if it cannot ultimately avoid, acknowledging errors to oneself and others. It also avoids being perceived as a "flip-flopper." It is not surprising that policy makers sometimes throw good money after bad by honoring sunk costs—for example, some members of Congress invoked the money already spent and lives lost in the NASA space shuttle program as a reason for continuing it (*Economist*, 2003). One might wonder whether the sunk-cost phenomenon underlies our "staying the course" in some international conflicts.

The best debiasing strategy here is to adopt the economist's or investor's perspective of disregarding costs incurred in the past and relentlessly focusing on the future. Doubtless easier said than done.

GROUP DYNAMICS

Irving Janis's classic study of *groupthink* (1972) was based on policy disasters: the failure to anticipate the Japanese attack on Pearl Harbor, the escalation of the war in Vietnam, the Bay of Pigs invasion, and the Watergate cover-up. Janis argued that groupthink is characterized by the group's premature convergence around a course of action without adequate analysis. He proposed that the phenomenon was a disorder of highly cohesive groups, exacerbated by ideological homogeneity, authoritarian leadership, and insulation from outside influences. Other studies of group dynamics have identified the *common knowledge effect* (Gigone and Hastie, 1993), where facts known to only one participant are unlikely to be shared with the group as a whole, and *group polarization* (Schkade, Sunstein, and Kahneman, 2000), where the members of even a heterogeneous group may converge on an extreme decision. Again, the Challenger Shuttle disaster manifested group decision-making pathologies (Vaughan, 1997).

These various disorders can be mitigated by employing a step-by-step process that specifies the problem and the interests at stake and considers a variety of solutions in terms of their benefits, costs, and risks before converging on a particular solution. The group must consciously avoid authoritarian leadership. It is helpful to assign "devil's advocates" to argue against an emerging consensus, and to poll individual members of the group both to elicit their unique knowledge and to obtain independent points of view. There is at least some evidence that such process improvements conduce to better outcomes (Schafer and Crichlow, 2002).

Coda: Policy Makers' Ignorance About Influencing Behavior

Skeptics of government regulation argue—not without some basis—that policy makers are often ignorant both about the effects of regulator interventions and about their own ignorance on the subject (Tasic, 2010). Whatever may be the general case, most policy makers are not aware of the range of factors that influence citizens' and consumers' behavior and of how policy interventions can leverage them. Hopefully, they will be informed by the essays in this volume and books such as Thaler and Sunstein's *Nudge: Improving Decisions About Health, Wealth, and Happiness* (2008) and Congdon, Kling, and Mullainathan's *Policy and Choice: Public Finance Through the Lens of Behavioral Economics* (2011).

Conclusion

It may turn out that some of the cognitive biases described in this book are hopelessly resistant to

debiasing and can only be met by strategies that counter one System 1 phenomenon with another, along the lines suggested by Weber. But the counterbiasing strategy presents the question posed in the title of this essay in potentially infinite regress, that is, who guards the guardians of the guardians?

My own preference is for true debiasing that corrects System 1 intuitions with System 2 analysis. Of the debiasing strategies discussed above, those that may affect a broad range of decisions include:

- Awareness of the biases
- Knowledge of probability, statistics and empirical methods
- Formal procedures that require considering opposite viewpoints and justifying one's conclusion

In truth, the research does not justify great optimism about debiasing in general (Fischhoff, 1975)—and least of all for legislators, who are democracy's most fundamental policy makers, but whose informal decision-making procedures make them highly vulnerable to cognitive errors, and whose accountability for outcomes can only amplify the biases of their constituents.

In the short run, I would put my hopes in courses such as the Woodrow Wilson School's "Psychology for Policy," which is required of the master's-degree students, and in executive education programs for policy makers, such as Jennifer Lerner's "Leadership Decision Making" at the Kennedy School. In the longer run, my hope lies in K–16 curricula that imbue citizens with critical thinking and problem-solving attitudes and skills.

Notes

I greatly appreciate comments on the draft by Iris Brest, Baruch Fischhoff, Lynne Henderson, Jennifer Lerner, Deborah Rhode, and Lee Ross. Much of this essay draws on Brest and Krieger (2010).

1. I do not address the interesting questions of law and political theory concerning whose utilities the CEOs of organizations or their agents and counselors should maximize.

2. This is true whether one views legislators as delegates, charged with representing their constituents' preferences, or as trustees, who follow their own understanding of the best action to pursue, See *political representation* in the *Stanford Encyclopedia of Philosophy* online (http://plato.stanford.edu/entries/political-representation/).

3. For example, transitivity entails that if you prefer having a red car to a blue car and a blue car to a green car, then you must prefer the red car to the green car. Procedural invariance entails that your preference for the color of the car cannot depend on the order that the salesperson shows the cars to you.

4. Gerd Gigerenzer and his colleagues argue that "fast and frugal heuristics" are equal, if not superior, to rational procedures in producing valid judgments of fact (e.g., Todd and the ABC Research Group, 1999). Without entering into this major debate, this essay generally endorses the value of rational procedures (Kelman, 2011).

5. Participants were asked to imagine preparing for the outbreak of a virus that was expected to kill 600 people. One group was asked whether to adopt Program A, which would save 200 people, or Program B, which had a one-third probability that 600 people would be saved and a two-thirds probability that no people would be saved. A second group was asked whether to adopt Program C, in which 400 people would die, or Program D, which had a one-third probability that nobody would die and a two-thirds probability that 600 people would die. Although Program A is identical to Program C and Program B is identical to Program D, participants tended to choose Programs A and D, demonstrating risk aversion when the problem was framed in terms of gains and risk taking when it was framed in terms of losses.

References

Ainslie G., and Haslam, N. (1992). Hyperbolic discounting. In G. Loewenstein and J. Elster (Eds.), *Choice over time* (pp. 57–92). New York: Russell Sage Foundation.

Alhakami, A. S., and Slovic, P. (1994). A psychological study of the inverse relationship between perceived risk and perceived benefit. *Risk Analysis*, *14*(6), 1085–1096.

Barlow, D. H. (1988). *Anxiety and its disorders: The nature and treatment of anxiety and panic*. New York: Guilford Press.

Blumenthal, J. A. (2007). Emotional paternalism. *Florida State University Law Review, 35*, 1–72.

Brest, P., and Krieger, L. H. (2010). *Problem solving, decision making, and professional judgment: A guide for lawyers and policy makers*. New York: Oxford University Press.

Brown v. Board of Education, 347 U.S. 483 (1954).

Burke, E. (1790). *Reflections on the revolution in France*.

Cialdini, R. B. (2008). *Influence: Science and practice* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Congdon, W. J., Kling, J., and Mullainathan, S. (2011). *Policy and choice: Public finance through the lens of behavioral economics*. Washington, DC: Brookings.

Daubert v. Merrell Dow Pharmaceuticals, 509 U.S. 579 (1993).

Dawes, R. M., Faust, D., and Meehl, P. E. (1982). Clinical versus actuarial judgment. In D. Kahneman, P. Slovic, and A. Tversky (Eds.), *Judgment under uncertainty:*

Heuristics and biases (pp. 306–334), New York: Cambridge University Press.

Ditto, P. H., and Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology*, *63*, 568–584.

Economist. (2003, August 28). The space shuttle: Old, unsafe, and costly. *Economist*, p. 77.

Federal Judicial Center. (2000). *Reference manual on scientific evidence* (2nd ed.). Washington, DC: Federal Judicial Center.

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, *1*, 288–299.

Fishkin, J. S. (2009). *When the people speak: deliberative democracy and public consultation*. New York: Oxford University Press.

Flyvbjerg, B., Bruzelius, N., and Rothengatter, W. (2003). *Megaprojects and risk: An anatomy of ambition*. Cambridge: Cambridge University Press.

Forgas, J. P. (1995). Mood and judgment: The affect infusion model (AIM). *Psychological Bulletin*, *117*, 39–66.

Giddens, A. (2008, January 2). This time it's personal. *Guardian*. Retrieved from http://www.guardian.co.uk/commentisfree/2008/jan/02/thistimeits personal

Gigone, D., and Hastie, R. (1993). The common knowledge effect: Information sharing and group judgment. *Journal of Personality and Social Psychology*, *65*, 59–74.

Guthrie, C. (2003a). Panacea or Pandora's box? The costs of options in negotiation. *Iowa Law Review*, *88*, 607–638.

———. (2003b). Prospect theory, risk preference and the law. *Northwestern University Law Review*, *97*, 1115–1163.

Guthrie, C., Rachlinski, J. J., and Wistrich, A. J. (2001). Inside the judicial mind. *Cornell Law Review*, *86*, 777–830.

Hammond, K. R. (2000). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice*. New York: Oxford University Press.

Hastorf, A. H., and Cantril, H. (1954). They saw a game: A case study. *Journal of Abnormal and Social Psychology*, *49*, 129–134.

Iyengar, S. S., and Lepper, M. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, *79*, 995–1006.

Janis, I. (1972). *Victims of groupthink: A psychological study of foreign policy decisions and fiascos*. Boston: Houghton Mifflin.

Jenni, K. E., and Loewenstein, G. F. (1997). Explaining the "identifiable victim effect." *Journal of Risk and Uncertainty*, *14*, 235–257.

Kahan, D. M., Slovic, P., Braman, D., and Gastil J. (2006). Fear of democracy: A cultural evaluation of Sunstein on risk. *Harvard Law Review*, *119*, 1071–1109.

Kahneman, D., and Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, and D. Kahneman (Eds.), *Heuristics and biases* (pp. 49–81). Cambridge: Cambridge University Press.

Kamin, K. A., and Rachlinski, J. J. (1995). Ex post ≠ ex ante: Determining liability in hindsight. *Law and Human Behavior*, *19*, 89–104.

Keeney, R. L. (1996). *Value-focused thinking: A path to creative decisionmaking*. Cambridge, MA: Harvard University Press.

Kelman, M. G. (2011). *The heuristics debate: Its nature and its implications for law and policy*. New York: Oxford University Press.

Kelman, M., Rottenstreich, Y., and Tversky, A. (1996). Context-dependence in legal decision making. *Journal of Legal Studies*, *25*, 287–318.

Keys, D. J., and Schwartz, B. (2007). "Leaky" rationality: How research on behavioral decision making challenges normative standards of rationality. *Perspectives on Psychological Science*, *2*, 162–180.

Klein, G. (2007, September). Performing a project premortem. *Harvard Business Review*, pp. 18–19.

Korobkin, R., and Guthrie, C. (1994). Opening offers and out-of-court settlements: A little moderation may not go a long way. *Ohio State Journal on Dispute Resolution*, *10*, 1–22.

———. (1997). Psychology, economics, and settlement: A new look at the role of the lawyer. *Texas Law Review*, *76*, 77–141.

Kunreuther, H., Novemsky, N., and Kahneman, D. (2001). Making low probabilities useful. *Journal of Risk and Uncertainty*, *23*,103–120.

Larrick, R. P. (2004). Debiasing. In D. Koehler and N. Harvey (Eds.), *Handbook of judgment and decision making* (pp. 316–338). Malden, MA: Blackwell.

Lehman, D. R., Lempert, R. O., and Nisbett, R. E. (1988). The effects of graduate training on reasoning. *American Psychologist*, *43*, 43l–443.

Lerner, J. (2012). Leadership decision making. Harvard Kennedy School, Cambridge, MA. Retrieved from http://ksgexecprogram.harvard.edu/Programs/ldm/overview.aspx

Lerner, J., and Shonk, K. (2010, September 1). How anger poisons decision making. *Harvard Business Review*, p. 606.

Lerner, J., and Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*, 255–275.

Loewenstein, G., and Lerner, J. (2003). The role of affect in decision making. In R. J. Dawson, K. R. Scherer, and H. H. Goldsmith (Eds.), *Handbook of affective*

science (pp. 619–642). Oxford: Oxford University Press.

Loftus, E. F. (1996). *Eyewitness testimony.* Cambridge, MA: Harvard University Press.

Lord, C. G., Ross, L., and 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.

Malouff, J., and Schutte, N. S. (1989). Shaping juror attitudes: Effects of requesting different damage amounts in personal injury trials. *Journal of Social Psychology*, *129*, 491–497.

McCleskey v. Kemp, 481 U.S. 279 (1987).

Milkman, K. M., Chugh, D., and Bazerman, M. H. (2009). How can decision making be improved? *Perspectives on Psychological Science*, *4*, 379–383.

Miller, D. T., and Taylor, B. R. (1995). Counterfactual thought, regret, and superstition: How to avoid kicking yourself. In N. J. Roese and J. M. Olson (Eds.), *What might have been: The social psychology of counterfactual thinking* (pp. 305–331). Hillsdale, NJ: Erlbaum.

New York Times. (2001, September 6). The statistical shark. Retrieved from http://www.nytimes.com/2001/09/06/opinion/the-statistical-shark.html

People v. Collins, 68 Cal.2d 319, 66 Cal Rptr. 497, 438 P.2d 33 (1968).

Rachlinski, J. J. (1996). Gains, losses, and the psychology of litigation. *Southern California Law Review*, *70*, 113–185.

Rawls, J. (1971). *A theory of justice.* Cambridge, MA: Harvard University Press.

Schacter, D. L. (2001). *The seven sins of memory: How the mind forgets and remembers.* New York: Houghton Mifflin.

Schafer, M., and Crichlow, S. (2002). The process-outcome connection in foreign policy decision making: A quantitative study building on groupthink. *International Studies Quarterly*, *46*, 45–68.

Schkade, D., Sunstein C., and Kahneman, D., (2000). Deliberating about dollars: The severity shift. *Columbia Law Review*, *100*, 1139–1175.

Shafir, E. (1993). Choosing versus rejecting: Why some options are both better and worse. *Memory and Cognition*, *21*, 546–556.

Siegel-Jacobs, K., and Yates, J. F. (1996). Effects of procedural and outcome accountability on judgment quality. *Organizational Behavior and Human Decision Processes*, *65*, 1–17.

Slovic, P., Finucane, M., Peters, E., and MacGregor, D. (2002). The affect heuristic. In T. Gilovich, D. Griffin, and D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 397–420). New York: Cambridge University Press.

Sunstein, C. R. (2005). *Laws of fear: Beyond the precautionary principle.* New York: Cambridge University Press.

Sunstein, C. R., and Kuran, T. (1999). Availability cascades and risk regulation. *Stanford Law Review*, *51* 683–768.

Sunstein, C. R., Schkade, D., Ellman, L. M., and Sawicki, A. (2006). *Are judges political? An empirical analysis of the federal judiciary.* Washington, DC: Brookings.

Tasic, S. (2010). *Are regulators rational?* Paper presented at the 7th Mises Seminar. Istituto Bruno Leoni, Sestri Levante, Italy. Retrieved from http://brunoleonimedia.servingfreedom.net/Mises2010/Papers/IBL_Mises2010_Tasic.pdf

Tetlock, P. E. (2000). Coping with trade-offs: Psychological constraints and political implications. In S. Lupia, M. McCubbins, and S. Popkin (Eds.), *Political reasoning and choice.* Berkeley, CA: University of California Press.

———. (2005). *Expert political judgment: How good is it? How can we know?* Princeton, NJ: Princeton University Press.

Tetlock, P. E., and Boettger, R. (1994). Accountability amplifies the status quo effect when change creates victims. *Journal of Behavioral Decision Making*, *7*, 1–23.

Thaler, R., and Sunstein, C. (2008). *Nudge: Improving decisions about health, wealth, and happiness.* New Haven, CT: Yale University Press.

Todd, P. M., and the ABC Research Group. (1999). *Simple heuristics that make us smart.* New York: Oxford University Press.

Tversky, A., and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453–458.

Twain, M. (1906–1907). *Chapters from my autobiography.* Project Gutenberg. Retrieved from http://www.gutenberg.org/files/19987/19987.txt

Vaughan, D. (1997). *The Challenger launch decision: Risky technology, culture, and deviance at NASA.* Chicago: University of Chicago Press.

Wilson, T. D., Centerbar, D. B., and Brekke, N. (2002). Mental contamination and the debiasing problem. In T. Gilovich, D. Griffin, and D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment.* Cambridge: Cambridge University Press.

Zimmerman, P., and Lerner, J. (2010, September 29). Decisions, decisions. *Government Executive.* Retrieved from http://www.govexec.com/dailyfed/0910/092910mm.htm

CHAPTER 30

Paternalism, Manipulation, Freedom, and the Good

JUDITH LICHTENBERG

The creature who has come to be known as homo economicus differs from living, breathing human beings in two central ways. First, homo economicus is fully rational: he always employs means that maximize the fulfillment of his ends and does what is in his best interests.[1] Human beings are often not rational; as a result of cognitive errors and biases, emotional reactions, and volitional weaknesses, they often fail to act in their own best interests. Behavioral economists and psychologists in this book and elsewhere have greatly increased our understanding of how human beings fall short in these respects and what can be done to more closely align their means with their ends (e.g. Barr, Mullainathan, and Shafir; Mullainathan and Shafir; Ubel; Wansink; all in this volume).

Human beings differ from homo economicus also in the ends they seek. Economists and others have often construed people's ends in terms of their narrow self-interest, particularly their economic self-interest. Yet—as careful thinkers note, although sometimes only when pushed—nothing in economic theory dictates the content of a person's ends or preferences. Assume an agent as altruistic as you please, whose deepest desire is to eliminate suffering and disease in the world. The fallacy of thinking agents must be self-interested results from confusing the *subject* of my preferences (me) and the *object* of my preferences (often me, but sometimes others) (Lichtenberg 2008, 2010a).[2] Several authors in this book acknowledge such other-regarding or altruistic preferences, which they call "social motivations" (Tyler, this volume; Weber, this volume). They suggest that we should capitalize on such motivations or try to enlarge their influence on behavior.

So behavioral economists and psychologists have called into question both assumptions about homo economicus—that he is rational (employs the optimal means to his ends) and that he is self-interested (cares only for his own well-being). But the two challenges pull in opposite directions. If human beings are less rational than homo economicus, then clearly they fall short. Their human traits constitute defects we should try to remedy or counteract, if we can do so without introducing other problems that are worse.[3] But if human beings are not (necessarily) self-interested, that is a good thing. Or so I shall assume. Homo economicus, then, is in one way worse and in another way better than real human beings.

Rationality and the Good

What does it mean to say that people sometimes act less than rationally? One way to understand the claim is to say that they fail to do what is in their best interests or to realize their own Good. For example, they fail to save adequately for retirement; they eat too much or unhealthily; they do not take their medicines as they should. But talking about a person's best interests immediately raises the question: best interests according to whom? In liberal societies it is natural to parse this concept in terms of an agent's own desires or preferences. In a common formulation, a person's best interests are what she would want if she possessed full information and suffered no cognitive, emotional, or volitional defects and biases. Such a definition might not always produce a determinate answer to the question of what is in a person's best interests, but we can suppose it does at least some of the time.

Of course, what people *would* want under these ideal and unrealizable circumstances is not equivalent to what they *do* want. That is part of the problem. But it may be less misleading to acknowledge that people have various wants and preferences that sometimes conflict. They want a comfortable retirement but also prefer more income now; they prefer to be fit and healthy but also like ice cream. Often such conflicts can be understood in terms of the distinction between short-term and long-term preferences. We can also distinguish levels or orders of preferences:

a smoker may have a first-order desire to smoke and a second-order desire not to smoke—that is, a desire not to desire to smoke. In any case, to say that people sometimes act less than rationally is to suggest that their desires can be ranked, preferably from their own point of view as well as from others'.

It is rarely helpful, however, to talk about what a person *really* wants, which suggests that although the person behaves as if she wanted one thing, in truth she wants something else. People's desires are multiple and conflicting. There is no plausible way to escape the conclusion that people hold inconsistent wants, desires, or preferences, and we should avoid linguistic tricks that seem to make these inconsistencies disappear.

Paternalism, Hard and Soft

Forcing people by law or some other form of regulation to act in their own best interests has traditionally been called *paternalism*. But several years ago Sunstein and Thaler introduced the concept of libertarian paternalism, which "attempts to influence the choices of affected parties in a way that will make choosers better off" without forcing them to do something or refrain from doing something (Sunstein and Thaler, 2003, p. 1162). So now we distinguish between classical and libertarian paternalism, or hard and soft paternalism.

It might seem almost a truism to say that if you can get people to change their behavior for their own good without forcing them, that is better than bringing the long arm of the law down on them. Yet although most people in liberal societies would prefer to avoid paternalism, there are probably irreducible differences in people's tolerance for it. Political libertarians think the price is always too high. Perhaps it is less misleading to say that they oppose it on principle. Others disagree; they think that the benefits of paternalism sometimes outweigh the costs.

Still, most people probably agree that we should minimize the use of coercion in guiding people to do what is good for them. We will be least uneasy if they choose freely and knowledgeably what is best for them. Alas, it turns out that information is not enough (Ubel, this volume). So the question is whether and to what extent we can induce (entice? cause?) people to do what is best for themselves—or, for that matter, others—without forcing them.

The validity of soft paternalism rests on at least two assumptions. One is that we can somehow formulate a coherent idea of a person's best interests, their Good—for example, in terms of what satisfies their long-term or higher-order preferences—and that it is better, other things being equal, if people achieve their Good than if they do not. It is not necessary that we be able to give a complete account of what is in a person's best interests, as long as we can give a determinate account in some cases.

The other assumption is that, as Thaler and Sunstein put it, there is no such thing as neutral design: every environment exhibits features—a "choice architecture"—that nudge agents in some direction rather than others, making it more likely that they will do X rather than Y or Z. A different way of putting the point is that human behavior is "heavily context dependent" (Barr, Mullainathan, and Shafir, this volume). In the psychological literature the technical term for this view is *situationism*, which insists on the power of situational factors over individuals' personal traits to determine behavior. Cafeteria items may be arranged in a variety of ways, but they must be arranged somehow, and their order may significantly influence people's food choices and thus their health (Thaler and Sunstein, 2008; Thaler, Sunstein, and Balz, this volume). Those who serve food must place it on plates of some size or other; plate size affects how much people eat (Wansink, this volume). Doctors must explain treatment options to their patients in some order, using particular language, and expressing probabilities in a particular way (McNeil et al., 1982; Ubel, this volume). Employers, governments, and others who offer policies regarding retirement benefits, insurance, organ donation, and other matters can offer opt-in or opt-out defaults (Johnson and Goldstein, 2003, this volume). These decisions may have profound effects on people's choices and thus on their well-being.

Paternalism and Manipulation

I want to make several points about the distinction between hard and soft (or traditional and libertarian) paternalism. First, as Thaler and Sunstein acknowledge (2008), the distinction is not sharp, since one can choose to violate even legally coercive rules, accepting the penalty or (more likely) taking the risk that one will not be caught. It does not follow that the distinction between legally coercive rules and other forms of influence is trivial, but we should note that influence is a matter of degree, with many points along the continuum between liberty and force.

Still, it is natural to think that not forcing people to act (or not act) is preferable to forcing them; better to leave the choice more open even if influence is inevitable. Yet in one way coercion might be preferable: it is overt and explicit. Citizens know that the state is attempting to control them when it prohibits riding

a motorcycle without a helmet. But they are likely not to notice the significance of the arrangement of food in the cafeteria or its influence on our behavior. Similarly with the default choice of retirement plans and other policies. The idea that someone is attempting to influence our choices without our knowledge or consent is troubling and may seem in some way at least as much a violation of our liberty as explicit coercion. We tend to call this kind of influence-creation *manipulation*; its connotations are negative.

One might respond that this objection neglects the idea that some arrangement or other of the choice environment is *inevitable* and that there is no neutral design. In this section I consider one aspect of this response; in the next section, another.

Suppose that nonneutrality is indeed inevitable. Still, manipulation might be reduced if policy makers were required to reveal more clearly how they attempt to influence decisions, so that agents could more easily resist their influence if they so chose. Of course, we know from behavioral economists and psychologists that awareness and knowledge are not always enough. Sometimes the difficulty is rather in the link between intentions (formed in light of knowledge)—with which the road to hell is paved—and action, as Barr, Mullainathan, and Shafir (this volume) argue.[4]

At the very least, designers of defaults can sometimes control how easy or hard it is to depart from them. For example, mortgage rules can be structured with opt-out defaults that "make it easier for borrowers to choose a standard product" and harder to choose one they are less likely to understand or to be able to afford (Barr, Mullainathan, and Shafir, this volume). Yet in many contexts transparency is unrealistic or impossible. Must the cafeteria managers explain the reason for their food arrangement or for the size of their plates? Must the Motor Vehicle Administration explain why it uses an opt-out rather than an opt-in default for organ donation? Transparency may be useful in some contexts but not in others.

Defaults

The second response to the claim that there is no neutral design is to question it outright. Consider the example of defaults, which seem to illustrate the nonneutrality thesis. Johnson and Goldstein (2003, this volume) have shown the profound effects of defaults on organ donation and other policies. Although organ donation is not a matter of paternalism but of other-regarding choices (about which I say more below), the mechanisms are the same as for paternalistic intervention.

In some countries, including the United States and Great Britain, you must choose (when you get or renew your driver's license) to become an organ donor; the default is not to donate. In many European countries, the policy is the reverse: consent to donating one's organs is presumed, and one must explicitly opt out to avoid donation. In Austria, France, Hungary, Poland, and Portugal, which all have opt-out policies, effective consent rates are over 99%. In countries with opt-in policies, consent rates are radically lower—from 4.25% in Denmark to 27.5% in the Netherlands.[5]

Yet a no-default policy is also possible: forced or mandated choice. In an online experiment, Johnson and Goldstein (2003) show that mandated choice approximates the opt-out default: 79% of participants who must decide choose to be organ donors; 82% in the opt-out default remain as donors; only 42% in the opt-in condition agree to be donors.

Are mandated choices counterexamples to the claim that neutral design is impossible? To fully answer this question would require an extended inquiry into the nature of neutrality, and even after it, we might still not reach a clear or uncontroversial answer. What seems certain is that mandated choice is *more* neutral than opt-in or opt-out defaults.

But that is not the end of the matter, because neutrality is not the only value and may not always be the most important one. Thaler, Sunstein, and Balz (this volume) argue that where choices are difficult or complicated, people may prefer a "good" or "sensible" default; and when choices are not binary, yes-no decisions, mandated choice might not even be feasible.

What is a good default? Perhaps it is the one I would prefer if I had full information and sufficient time and mental resources to process it. Since people have different values and preferences, on this criterion no default is necessarily best for everyone. Some people would like to donate their organs, but some object on religious grounds. So the good default might be the one that most people would prefer. In the case of organ donation, Johnson and Goldstein's online experiment suggests that opt-out policies are better because they more closely match people's preferences when no default is offered. Even apart from cases where choices are not binary, however, it is implausible to think that people have preexisting preferences in many situations in which defaults are common and desirable (Thaler and Sunstein, 2003, pp. 1173–1174). I may not have a preference concerning the details of my software installation, even armed with full information and adequate mental resources. More serious still is that our preferences are partly constructed out of the choice situations in which we find ourselves and thus cannot be employed to structure those choice situations.

What can we conclude from this discussion? First, even if we agree that there is no neutral design, some

designs may be more neutral than others. But, second, neutrality does not always trump all other values. Especially if the aim is to do what is in people's best interests or satisfy their (deeper? more important? more enduring?) preferences, we will sometimes want to structure environments in ways that are in tension with choices they might otherwise make.

Politics, Power, and Freedom

One of Thaler and Sunstein's central aims seems to be to reassure those who worry about bringing the state's power down on individuals through paternalistic legislation that such crude techniques are not necessary. But the message of their work, and that of other behavioral economists and psychologists, might be seen in less rosy terms, a glass half empty rather than half full. Despite the desire to preserve freedom that leads us to resist hard paternalism, we are not very free at all. Subject to error, bias, ignorance, temptation, passion, and weakness of will, we find ourselves (or, more often, fail to realize that we are) buffeted about by the winds of influence, internal and external, intentional and accidental, self-interested and benevolent. We can learn to control some of the forces acting upon us so that we are better able to realize our Good, but sometimes it may seem not much more than a rhetorical trick to say we are thereby free.

Despite the wealth of insights behavioral economists and psychologists have provided, with a few exceptions there is a peculiarly apolitical quality to their work. One might infer from the literature that the cognitive, affective, and volitional deficiencies that lead agents astray are merely unfortunate natural facts; one might fail to see how they are actively exploited and encouraged by banks, insurance and credit card companies, fast-food conglomerates, and others who profit from these weaknesses. Altering the choice architecture so as to nudge people to serve their own best interests is important. But some entities need more than nudges. The activities of corporations and others who prey on individuals need nonpaternalistic, other-regarding restrictions, in addition to positive requirements that they serve individuals' interests (for important examples see e.g., Barr, Mullainathan, and Shafir, this volume; Mullainathan and Shafir, this volume). This is less a matter of behavioral economics in the usual sense than of the realities of politics and power.

Rationality and Morality

As I noted at the outset, economists and other social scientists often shrink from assuming that people are capable of acting altruistically. That reluctance may derive from a belief in egoism as a kind of default—the uncontroversial view that needs no defense and that keeps social science away from the dangerous territory of "value judgments." Yet the clear implication of behavioral economics and psychology (not to mention philosophy) is that we cannot avoid making value judgments. If there is no neutral design of choice environments, or if even the choice of a neutral design is itself not neutral (as the discussion above of reasons against no-default choices suggests), we have no alternative but to shape choice environments in accordance with some values or other. We should do so in accordance with a conception of what is genuinely in people's best interests or which preferences it is most important for them to satisfy. To leave the environment as it is (whatever that might mean) is also to make a value judgment, and the jumble of people's conflicting desires and preferences forces us to favor some and not others.

From the recognition that we need a conception of a person's good it is not much of a step to the conclusion that we need a conception of the general good. The value judgments inherent in the general conception are no more significant than in the individual, the gap between my immediate preferences and my best interests no wider than the gap between my good and your good.[6] Two other facts lend support to the legitimacy of taking into account more than people's egoistic choices. One is that, as others in this volume have argued (Tyler; Weber), individuals have social motivations: they care not only about themselves but also about others. In other words, they are somewhat altruistic (some more than others, of course).

The other is that nudging individuals to act in accordance with the interests of other people is rooted not only in the assumption that they would so choose but also in the fact that they have moral responsibilities to do so. The more minimal defense of these responsibilities rests on so-called negative duties. When our actions contribute to harming other people (or creating "externalities," as economists like to put it), those harmed may have valid claims against us; in many such cases the state is entitled or perhaps even required to enforce such claims. This much even political libertarians admit! Attempts to induce people to behave in ways less harmful to the environment can be rooted in these negative duties. Somewhat more controversial is the idea that we have not only negative duties not to harm others but also, at least sometimes, positive, "humanitarian" duties to help them. But how controversial is this view really? Do we need fancy arguments to be convinced that it would be better if people did not ignore genocide and other atrocities and that it is therefore legitimate to shape environments in ways that cause them to act accordingly (Slovic et al., this volume)?[7]

Notes

1. Whether these are equivalent is an open question I address briefly in what follows.

2. For experimental evidence of unselfish motives see, e.g., Batson (1991); Fehr and Fischbacher (2004). For most purposes the existence of unselfish motives is pretty obvious, but at a deep level, the claim is difficult to test, as Batson acknowledges. He and his colleagues attempt to test it through a number of complex experiments, all of which confirm the existence of altruistic motivations. As Sober and Wilson note (1998), however, this does not prove that all versions of egoism will fail. Because sophisticated versions appeal to the internal rewards of helping others—rather than simply money, say—it is always possible that a more subtle psychological reward lurks that the experiments have not detected (pp. 271–273). This possibility will strike many as far-fetched, confirming their suspicions that egoism is unfalsifiable, but it permits those attracted to egoism to hang on to their convictions.

3. Perhaps not all such differences between homo economicus and real human beings should be construed as defects. I leave that question aside, assuming only that at least some of these traits are flaws.

4. They discuss changes to the Truth in Lending Act that require credit card companies to disclose to customers information about the expected time it will take to pay off credit card balances if they pay only the minimum balance, and they argue that "such disclosures may not be strong enough to matter. . . . In fact, the borrower would need to change behavior in the face of strong inertia and marketing by credit card companies propelling her to make no more than the minimum payments."

5. Johnson and Goldstein offer three (non–mutually exclusive) explanations for the power of defaults: effort, implied endorsement, and loss aversion. Sunstein and Thaler (2003) suggest another important one: the idea that the default is "what most people do, or what informed people do" (p. 1180). This might appear similar to implied endorsement. But there are two possible differences. First, Johnson and Goldstein's idea focuses on the policy maker's endorsement, Sunstein and Thaler's on the public's. Second, an agent may choose what she believes is the popular choice not because people's choosing it signifies approval of some independently valuable good but simply because she wants to do what others are doing, irrespective of whether it has independent merit.

6. For a view showing the similarities between prudential and moral reasons see Nagel (1970).

7. For an argument that the distinction between negative and positive duties—between the duty not to harm and the duty to render aid—is exaggerated, see Lichtenberg (2010b).

References

Batson, D. (1991). *The altruism question: Toward a social-psychological answer.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Fehr, E., and Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences*, *8*(4), 185–190.

Johnson, E. J., and Goldstein, D. G. (2003). Do defaults save lives? *Science*, *302*(5649), 1338–1339.

Lichtenberg, J. (2008). About altruism. *Philosophy and Public Policy Quarterly*, *28*(1–2), 2–6.

———. (2010a, October 19). Is pure altruism possible? *New York Times.* Retrieved from http://opinionator.blogs.nytimes.com/2010/10/19/is-pure-altruism-possible/

———. (2010b). Negative duties, positive duties, and the "new harms." *Ethics*, *120*(3), 557–578.

McNeil, B. J., Pauker, S. G., Sox, Jr., H. C., and Tversky, A. (1982). On the elicitation of preferences for alternative therapies. *New England Journal of Medicine*, *306*(1), 1259–1262.

Nagel, T. (1970). *The possibility of altruism.* New York: Oxford University Press.

Sober, E., and Wilson, D. S. (1998). *Unto others: The evolution and psychology of unselfish behavior.* Cambridge, MA: Harvard University Press.

Sunstein, C. R., and Thaler, R. H. (2003). Libertarian paternalism is not an oxymoron. *University of Chicago Law Review*, *70*(4), 1159–1202.

Thaler, R., and Sunstein, C. (2008). *Nudge: Improving decisions about health, wealth and happiness.* New Haven, CT: Yale University Press.

Index

9/11 attacks, 232
9/11 Commission Report (National Commission on Terrorist Attacks Upon the United States), 138
401(k) plans, 444; automatic enrollment in, 465; defaults in, 468
1540 Committee, 139
1984 (Orwell), 109
1/n rule, 245, 261n2

Aaker, J. L., 284
Abadie, A., 418–419
Abu Ghraib, 137
accountability movement, 330
Accreditation Council on Graduate Medical Education, 41
adjudicative fact-finding, 484; adjudicator procedures, 484–485; and percipient witnesses, 484; and reliance on clinical rather than on statistical prediction, 485
Adkins, N. R., 284
advertising: and approach motivation, 302–303; Don't Mess with Texas campaign, 303, 304; Friends Don't Let Friends Drive Drunk campaign, 303–304; social-norms marketing, 304
affect, 127–128; affective imagery, 137–138; experience of, 126–127; and System 1 and System 2 approaches, 127, 128, 133. *See also* affect, analysis and the value of human lives
affect, analysis and the value of human lives, 128; and the psychophysical model, 128–130
African Americans, 22, 23, 34, 35, 41, 44, 274, 329, 331; and GPA scores, 336–337. *See also* racial bias; racism; White Castle, killing of an African American by police at
age/aging: age-discrimination law, 264, 269; anti-elder bias, 23; stereotyping of, 17
Age Discrimination in Employment Act (ADEA [1978]), 269
AIDS, 240
Akerlof, G. A., 270, 271, 277n2, 278n9
Al Qaeda, 308
Alar, 233
Allais paradox, 385
Allport, G. W., 34, 40, 42, 45
Alter, A. L., 182
Ambady, N., 275, 284
ambiguity aversion, 385
American Association of University Women (AAUW), 53
American Judicature Society, 155
American Medical Association (AMA), 40
American Payroll Association, 285, 443
American Psychology Law Society, 172
Ameriks, J., 255, 257–258
Ames, D. R., 390–391
analysis: descriptive analysis, 218; normative analysis, 218; prescriptive analysis, 218
Anderson, N., 121
anecdotes: balanced anecdotes, 357; statistically reinforced anecdotes, 357
Angel, David, 157
ANOVA test, 258
Arab-Israeli conflict, 112, 121
arbitration agreements: confidentiality agreements, 65; mandatory agreements, 65; predispute agreements, 65; protective orders, 65
Arendt, Hannah, 23
Aron, A., 38
Aronson, Joshua, 332, 335, 340
Arthur, W., Jr., 44
Asch, Solomon, 23, 109
Ashmore, R. D., 18, 22
Asian Americans, 18, 37–38, 331
Aspen Institute Congressional Program, 487
atrocity. *See* psychic numbing, and mass atrocity
attitudes: explicit attitudes, 34; implicit attitudes, 34
"audit studies," 43
Austin, J. L., 183
awareness/education, as solutions to overeating, 315–316; meta-analysis of awareness, 316, 318; and the problem of awareness, 316; why education and awareness do not seem to work, 318
Ayres, Ian, 458n8

Bailis, D., 232
Bandura, Al, 331
Barberis, N., 257
Barlow, D. H., 488
Baron, J., 357
Barr, M. S., 447, 453, 458n5
Barron, L. G., 43
Bartel, C. A., 44
Bartels, D. M., 388
Batson, Daniel, 2, 127, 133, 282, 441, 498n2
Bazerman, M. H., 201, 209
Beaver, Antonio, 167–168
Becker, G. S., 186, 425
Bedau, H. A., 163
behavior, 440; behavioral interference paradigm, 16; and context dependence, 2, 440–441; and decisional conflict, 441–442; individual behavior, 22; knowledge and attention, 442; mental accounting, 442; and the power of institutions, 442–443. *See also* behavior change
behavior change, 301; approach and avoidance motivation, 302–303; levers of behavior change, 303–304; psychological approaches (attitude-change approach and motivational dynamics approach) to behavior change, 301–302
behavioral field economics, 469; behavioral public finance (diagnosis, judgment, prescription), 469–471; other fields, 471
behavioral mimicry, 44
behaviorally informed regulation, 440, 457; behavior, markets, and policy, 444–446; behaviorally informed credit card regulation, 453–455; behaviorally informed home mortgage regulation, 447–453; increasing saving among low- and moderate-income (LMI)

behaviorally informed regulation (*cont.*)
households, 455–457. *See also* behavior
Behrman, B. W., 154
Benabou, R., 375n1
Benartzi, S., 245, 246, 253, 261n2, 265, 268, 367, 374, 383, 393
Bentham, Jeremy, 186, 187, 188
Bernheim, B. D., 255, 277n1
Bernoulli, D., 385
Bernstein, Penny, 150
Berry, C., 291
Bertrand, M., 441–442
Bewley, T. F., 270
bias, 195, 212, 481; action and inaction biases, 489; alarmist bias, 234; anchoring and insufficient adjustment, 485; anti-elder bias, 23; and behavior, 206; "bias blind spot," 483; biased assimilation, 109–110, 208; biased perception and perceptions of bias, 111–112; biases in social perception stereotyping, 487; confirmation bias, 171, 175, 478, 485, 487; and counterbiasing, 483–484; flat-rate bias, 370; hindsight bias, 217, 484–485; implicit bias, 274, 275–276; in-group bias, 61; judgment of debiasing efforts, 356; and knowledge, 352–353; outcome bias, 217; perceptions of in conflict and negotiation, 108–110; projection bias, 365, 404; relevance of to public policy, 195–196; risk of bias, 212; self-serving fairness bias, 364; single-action bias, 383, 392; status-quo bias, 249–250, 363; witnesses' biases, 484. *See also* bias, causes of asymmetry in; bias, perceptions of in the self and others; bias, policy application case studies; disclosure, mandating of; gender bias; objectivity; perspective taking
bias, causes of asymmetry in, 199; disagreement and naive realism, 200; self-enhancement and the motive to deny bias, 201; unconscious bias and introspection illusion, 199–200
bias, perceptions of in the self and others, 196; cognitive biases, 199; correspondence bias, 199; and the "planning fallacy," 199; prejudice and group-based biases, 198–199; self-enhancement biases, 196–198; self-interest biases, 198
bias, policy application case studies, 201; conflict, 203–204; ethical lapses, 201–202; persistence of racism and sexism, 202–203
Blais, A. R., 386, 390–391
Blanco, Kathleen, 407–408
Blinder, A. S., 270
Blink (Gladwell), 475
blood donation, 306
Bloodsworth, Kirk, 150
Blount, S., 44
Blue Zones Vitality Project, 325
Bodie, Z., 257, 262n19
Borchard, Edwin, 163
Boston Research Group, 251
bounded rationality, 6, 274, 265, 274, 277, 380, 386, 478
bounded self-interest, 6, 264, 270, 277, 478
bounded willpower, 6, 264, 265–266, 267–269, 270, 277, 277n2, 277–278n4, 478
Brann, P., 86
Briley, R. M., 284
Bronzaft, A. L., 292
Brown, C. L., 421
Brown v. Board of Education (1954), 32, 34, 486
Bruner, Jerome, 109, 341
Bruzelius, N., 485
Bryan, C. J., 109
Burby, R. J., 399, 403, 407, 408
Burgess, D., 42
Burke, Edmund, 485
Burlingame, L., 451
Bush, George W., 110–111, 126
Butsch, S., 204

Cain, D. M., 206
cake mixes, 304–305
California Water Code, 185
Caligiuri, H. L., 154
Campbell, C. M., III, 270
Canada, 232
Canadian Standards Association, 227
Cannino, Jennifer Thompson, 150
Cantril, H., 112
Capital Punishment Reform Study Committee, 153
Capozzi, Anthony, 145
carbon footprint, 388–389, 454–455
CARD Act (2009), 445, 447, 453
Carlin, George, 114
Carlsmith, K. M., 183, 184
Carraher, D. W., 341
Caruso, E., 209
Center for Drug Evaluation and Research (CDER), 221
Center for Economic Progress, 289
Center for Modern Forensics Practice, 152, 155
Center for Nutrition Policy and Promotion (CNPP), 324
certainty effect, 385
Chandon, P., 313
Charles Schwab, 256–257
Cheney v. United States District Court for the District of Columbia (2004), 207
China, 128
Choi, D. H., 270
Choi, J. J., 246, 250, 261n8
choice architecture, 245–246, 261, 428–430, 466; and asymmetric paternalism, 467–468; benefits and limitations of, 468–469; expecting error, 431–433; giving feedback, 433–434; and incentives, 437–438; and nudges, 466–467; structuring complex choices, 435–437; understanding mapping, 434–435. *See also* choice architecture, and escalator programs; choice architecture, and portfolio choices; defaults
choice architecture, and escalator programs: background of, 247; and hyperbolic discounting, 247–249, 252, 266; and inertia (the "status-quo bias"), 249–250; and nominal loss aversion, 250–252
choice architecture, and portfolio choices: asset allocation funds and equity market participation, 255–257; asset allocation funds as the default investment option, 262n18; background of (risk-based funds, retirement date finds, asset allocation funds), 252–253; and the "core plus" strategy, 255; data and descriptive statistics concerning, 253–255; and the narrow-framing tendency, 257; and the "participation gap," 255; and plan selection bias, 258–269; and the "psychic costs argument," 257; retirement date funds and lifecycle investment patterns, 257–261, 262n20
choice bracketing, 364–365; bracketing effects, 364–365; in weight-loss programs, 370
Cialdini, R. B., 98
Civil Rights Act (1964), 32; Title VII of, 42
Civil Rights era, 185
Claire, T., 284
Clark, Maurice, 1
climate change, 236, 381, 383–383, 388, 392, 393, 485, 487
Clinton, Bill, 126
Coffee, J. C., Jr., 185
cognition, 292, 294; cognitive load, 292; cognitive operations (System

I and System II), 231, 274; and stress, 293
cognitive psychology, 281
Cohen, G. L., 198–199, 211, 333, 336, 338
Cole, M., 341
collaborative filtering, 436
Colom, Alvaro, 190
Commission on Capital Punishment, 152
Common Ingroup Identity Model, 39, 42
common knowledge effect, 490
competence, and public policy, 217; approach to competency-based policy making (normative, descriptive, and prescriptive analysis), 217–218; and blanket claims of competence, 227; and defensive attribution, 217; importance of an interdisciplinary team (subject-matter experts, decision analysts, social scientists, and designers) to, 227; organizing for decision-making competence, 227–228; strategies for competence assessment (reprise of the case studies), 226–227
competence, and public policy, case study of (carotid endarterechtomy), 227; descriptive analysis of, 223; normative analysis of, 223; policy analysis of, 223–224; policy context of, 223; prescriptive analysis of, 223–224
competence, and public policy, case study of (*Cryptosporidium* contamination emergency), 227; descriptive analysis of, 225; normative analysis of, 225; policy analysis of, 255; policy context of, 225; prescriptive analysis of, 225
competence, and public policy, case study of (emergency evacuation/RRD [radioactive dispersion devices] scenario), 227; descriptive analysis, 226; normative analysis, 226; policy analysis of, 226; policy context of, 225–226; prescriptive analysis of, 226
competence, and public policy, case study of (methylene chloride-based paint stripper), 227; descriptive analysis of, 224; normative analysis of, 224; policy analysis of, 224–225; policy context of, 224; prescriptive analysis of, 224. *See also* competence, and public policy, case studies
competence, and public policy, case study of (Plan B morning-after pill), 221, 227; descriptive analysis of, 222; normative analysis of, 221–222; policy analysis of, 222–223; policy context of, 221; prescriptive analysis of, 222
competence, and public policy, case study of (saw palmetto), 226–227; descriptive analysis of, 220; normative analysis of, 219–220; policy analysis of, 220–221; policy context of, 219; prescriptive analysis of, 220
computer simulations, 392
conditional cash transfer (CCT) programs, 307
conflict, 353; decisional conflict, 352. *See also* conflict resolution
conflict resolution, 108; and the inevitability of resolution, 121. *See also* conflict resolution, barriers to; conflict resolution, methods for employing
conflict resolution, barriers to, 115; acknowledging negotiation expectations and ideology, 121–122; cognitive dissonance and rationalization, 115–116; concession and reactive devaluation, 118; and education, 122; management of attributions, 120–121; and the pursuit of equity and justice, 116–117; reactive devaluation, loss aversion, and reluctance to trade, 117–118; and third-party mediators, 116, 120
conflict resolution, methods for employing, 118; countering false polarization, 118–119; framing and construal, 119–120; multitrack diplomacy, 119; public peace processes, 119
Congdon, W. J., 490
conjoint analysis, 354–355
consequentialism, 238
conservative lag, 305
consumption, stimulation of, 312; salience of food, 312; and the shape of serving containers, 313–314; size of food packaging, 313; stockpiled food, 313; structure and variety of food assortments, 312–313
consumption, subsidizing of, 307–308
consumption, volume of, 310–311; monitoring of consumption volume, 311–312; and the power of consumptive norms, 311
construal, 282; "bottom-up" construal of issues, 112; construal differences, 113; and framing, 119–120; the role of "construal" in mental life, 2
Construal Level Theory, 189
Consumer Financial Protection Bureau (CFPB), 447, 448, 452
Cooper, J., 43
cooperation, 77, 84–85, 88; framework for voluntary cooperation, 78–79; importance of identification to, 87–88; in the Lewinian tradition, 88; with the police, 88n3. *See also* social motivations
Cormier, Warren, 251
Cornell University Food and Brand Lab, 325
cost-benefit analysis, 231, 240, 241n8, 383; eight propositions concerning, 239–240; as political versus metaphysical, 238–239; problems with aggregated willingness to pay, 238. *See also* cost-benefit analysis, and correction of misfearing; cost-benefit analysis, objections to
cost-benefit analysis, and correction of misfearing, 231–232, 235–236, 383, 477; and aggravating social influences (information and reputational cascades), 233, 240; and the availability heuristic, 232–233, 240; emotions and probability neglect, 233–234; general implications of, 235; systemic effects and "health-health" trade-offs, 234–235
cost-benefit analysis, objections to, 236; populism, 236; qualitative differences among social goods, 236–237; rival rationalities, 237–238
counselors, 481–482
Coursey, D., 403
Crane, D. B., 257, 262n19
criminal investigation and adjudication, structure of, 168; arrest and charging of the criminal, 168–169; identification of the criminal, 168; pretrial sorting, 169; review and appeal, 169–170; trial, 169
Croson, R., 374
Crozier, J. C., 284
Cruz, Rolando, 170–171
cue-utilization theory, 293
Cushman, F., 182

Dana, J., 202
Darley, John, 2, 182, 183, 184, 188, 282, 441
Darling, S., 154
Daubert v. Merrell Dow Pharmaceuticals (1993), 484, 486
Davey, S. L., 154
death penalty, 109–110, 113–114
debiasing, 483–484, 486, 487
decategorization, 38, 39

deception, 108; self-deception, 476
decision aids (DAs), 351, 357, 358; and the assessment of aggregate responses, 356; health-care decision aids, 352; and the inadequacy of decision-making criteria, 353–353
decision errors, 361, 374–375; as instances of misweighting, 363; as instances of reweighting, 363; and the theory of the second best, 362–363. *See also* decision errors, applications of behavioral economics at the individual level; decision errors, applications of behavioral economics at the societal level; decision errors, specific, and behavior improvement
decision errors, applications of behavioral economics at the individual level: medication (warfarin) adherence, 371–372; saving, 365–368; weight loss, 368–370
decision errors, applications of behavior economics at the societal level: charitable giving, 373–374; global warming, 372–373; international disputes, 374
decision errors, specific, and behavior improvement: loss aversion, 363; narrow bracketing, 364–365; nonlinear probability weighting, 364; overoptimism, 365; peanuts effects, 264; present-biased preferences, 363–364; projection bias and hot-cold empathy gaps, 365; the self-serving fairness bias, 364; the status quo, or "default," bias, 363
decision making: affect-based decision making, 389–390; avoiding difficult trade-offs, 489; calculation-based decision making, 389; case-based decision making, 390; and choice overload, 488; and cognitive myopia and excessive discounting of future benefits, 489; and context dependence, 488; decisions from description, 386; decisions from experience, 386; defining the problem and considering solutions, 488; and the effect of the decision maker's emotional state, 488–489; and identity salience, 283–284; and loss aversion, behavior under risk, and action and inaction biases, 489; passive decision-making, 293; preference-sensitive decision making, 351; reason-based decision making, 488; recognition-based decision making, 390; role-based decision making, 390; rule-based decision making, 390, 392; sequential consideration of choices, 488; value-based decision making, 488. *See also* decision aids (DAs); decisions, criteria for judging
decisions, criteria for judging, 353; adherence, 358; conjoint analysis, 354–355; correlation validity, 357–358; the happiness criterion, 356; the invariance criterion, 356–357; and the reduction of mispredictions, 355–356; standard gamble method, 354; time-processing the decision, 358; and the time-trade-off method, 354; the unexpected utility criterion, 353–355
Declining Significance of Gender, The (Blau, Brinton, and Grusky), 52
default case studies: Do Not Call Registry, 420; insurance decisions, 417; Internet privacy policies, 419; military recruitment and the No Child Left Behind Act, 419–420; organ donation, 417–418; retirement savings, 419; sex education in Kansas, 419
default effects, causes of: effort, 420–421; empirical evidence, 422; implied endorsement, 421; loss aversion and reference dependence, 421–422
defaults, 8–9, 252, 262n18, 393, 417, 430–431, 467, 496–497; benign defaults, 423–424; costs, benefits, and efficient defaults, 424–435; default options, 417; effects of on organ donation, 261n10; managing defaults, 423; mass defaults, 423; names of, 425n1; no-action defaults, 417; persistent defaults, 424; personalized defaults, 424; random defaults, 424; smart defaults, 424. *See also* default case studies; default effects, causes of
defensive attribution, 217
defined benefit plans, 252, 366
Dehaene, S., 129
DeParle, J., 289
Derhei, J. V., 257
Design of Everyday Things, The (Norman), 429
deterrence: behavioral perspective (duration neglect), 189; behavioral perspective (time discounting and hyperbolic discounting), 188–189; effectiveness of increases in sentence duration, 187; by increasing sentence duration, 185–186; and probabilistic future outcomes, 187–188; salience-based system of, 189–190; and "three strikes and you're out" laws, 187; in the United States, 186–187
devaluation, 335; group-based devaluation, 334; reactive devaluation, 4, 108, 117, 118, 120, 122
Devine, Patricia, 14–15, 16
Diamond, S. S., 155
Dichter, Ernest, 302
Diekman, K., 116, 117
Dillard, Annie, 128, 132
Dinner, I., 422
disclosure, mandating of, 204; problems associated with, 204–206
disclosure, strategies for implementing, 206; accept the need for more than disclosure, 207–208; educate disclosers, 206; educate recipients, 207; introduce disclosure requirements, 206–207; make disclosures readable and transparent, 207; use of third-party disclosers, 207
discounting, 6, 189, 266, 385, 391, 402, 421; asymmetric discounting, 266; excessive discounting, 7, 236, 392, 421; hyperbolic discounting, 188, 247–249, 252, 266, 363–364, 401, 402, 404, 409, 472; intertemporal, 385; temporal/time discounting, 188, 285, 363, 367, 388, 401, 443, 476
discrimination, 4, 5, 6, 23, 33, 34, 45, 52–53, 59–60, 61, 64, 67n1, 201, 202, 293, 332, 333; age discrimination, 264, 269; employment discrimination, 265, 273, 274, 275, 277; forms of subtle discrimination, 477; gender/sex discrimination, 42, 65, 66, 211; identification of, 59, 66; intergroup discrimination, 14; latent discrimination, 478; racial discrimination, 14, 35, 39
disidentification, 334
dispute resolution. *See* conflict resolution
diversity training, 42, 44–46, 64
Do-Not-Call Improvement Act (2007), 426n4
Do Not Call Registry, 420, 421
Dobbin, F., 64
Doctors for Life International court case, 139
Dodd-Frank Wall Street and Consumer Protection Act (2010), 447, 450, 452–453, 457
Dodson, Gary, 177n1
Dombroski, M., 226
Donate Life Illinois, 438n4
Donohue, John, 60

Don't Mess with Texas campaign, 303, 304, 308
Doob, A. N., 187
"door-in-the-face" effect, 374
Dorner, Dietrich, 234
double blinds, 146, 155, 157–158n6, 212, 335; double-blind administration, 149, 150–151; double-blind lineups, 146, 148, 149; sequential double-blind lineups, 149–150, 151, 152, 153, 157n5
Douglas, D., 458n19
Doverspike, D., 44
Dovidio, J. F., 35, 37, 43
Doyle, James, 152, 153
Drizin, S. A., 169
"dual-process theories" of thinking, 127
Dubner, Stephen, 475
dueling, in the South, 305
duration neglect, 404
Dweek, Carol, 331
Dwyer, P. D., 253

Earned Income Tax Credit (EITC), 284
Easterbrook, J. A., 293
eating behavior, 310, 314, 325–326; macro- and micro-level analysis of, 318; suggestions for changing eating habits (provide evidence that change will work, provide encouragement and feedback, provide a stylized set of changes; provide a tool for personal accountability), 321–322; turning mindless eating into healthy eating (*See also* Mindless Eating Challenge), 318–319. *See also* awareness/education, as solutions to overeating; consumption, stimulation of; consumption, volume of; eating behavior, partnerships for changing
eating behavior, partnerships for changing, 322; case study, 322–323; government partnerships, 323–324; need for partnerships, 325; state extension and frontline nutrition, 324–325. *See also* Center for Nutrition Policy and Promotion (CNPP)
economic/psychological incentives, nonadditive effects of, 304; calibration of taxes and subsidies, 308; counterintuitive effects of economic subsidies, 306–308; counterintuitive effects of economic taxes, 304–306
economics: behavioral economics, 362; behavioral field economics, 469; interface of with psychology, 78; standard economics, 361
Edin, K., 286
education, 122; cross-cultural education, 41; multicultural training education programs, 45. *See also* educational intervention
educational intervention, 329, 342–343; and attributional retraining, 335; and the classroom as a social tension system, 329–331; and identity threat, 332–333; and lessening the impact of social identity threat at the appraisal stage, 336–338; and the minority achievement gap, 331–333; and socioeconomic status (SES), 331. *See also* educational intervention, general lessons concerning; identity engagement process
educational intervention, general lessons concerning, 338; importance of timing, 341–342; no intervention is an island, 339; one size does not fit all, 341; prepare for the intervention and do not oversimplify, 339–340; psychological factors involved in intervention, 340–341; small things matter, 338–339
egalitarian societies, 22
egalitarianism, 23, 33
Eggers, S. L., 219, 221, 222
"ego depletion," 293, 321
Eichmann, Adolf, 131
"elephant effect," 485
Elias, J. J., 425
elimination by aspects, 436
Ellsberg, D., 385
Ellsberg effect, 385
Ellwood, C. A., 390
Employee Retirement Income Security Act (ERISA [1974]), 264, 267–268, 277–278n4
employment law, 264–265, 277; and age-discrimination law, 264, 269; and "attribute substitution," 274; and back-loaded wages, 269; bias-reducing effects of employment discrimination law, 275; and bounded willpower, 264, 265–266, 267–268, 277n2; and the costs of the minimum wage law, 273; and departures from expected utility theory and employment mandates, 276–277; and the efficiency wage model, 270; and the exemption of domestic service employees, 271–272; and the failure to cover independent contractors, 272–273; and fairness dynamics, 264–265, 270–271, 273–274; implicit bias and employment discrimination law, 275; and judgment errors, 265, 274; limits of existing employment discrimination law, 275–276; and pension regulation, 267–268; and social security, 264, 268–269; utility theory and employment mandates, 276; wage payment law, 266–267
Endangered Species Act (1973), 241n9
endowment effect, 265, 276–277, 278n18, 283
Engel, C., 390
environmental decisions, 380–381, 391–393. *See also* environmental decisions, behavioral insights; environmental decisions, behaviors of concern
environmental decisions, behavioral insights: decision-making unit, 388; gain versus loss framing and risk and loss aversion, 387; mental accounting, 388–389; multiple modes of making decisions, 389–391; multiplicity and flexibility of goals, 389; social comparisons and regret, 387–388
environmental decisions, behaviors of concern, 381–382; cognitive myopia, 383; effects of small probabilities, 386–387; hyperbolic discounting, 384–385; insufficient visceral reactions to environmental risks, 382–383; loss aversion, 383–384; risk and ambiguity aversion, 385–386
environmental externalities, 474–473
Epley, N., 209
Epstein, S., 127
Equal Employment Opportunity Commission (EEOC), 61, 64; organizational compliance with, 64–65
equal employment opportunity (EEO) policy, 60; information asymmetries as a source of inefficiency in EEO compliance, 66–67
ethnic diversity, 2
European Americans, 329, 331, 336
exercise. *See* physical exercise
expected utility theory, 264, 265, 274, 276, 383–384, 385, 400, 401
eyewitness identification, and the legal system, 145, 157; and the constructive memory process, 146; double-blind studies of, 146; eyewitness error, 166; eyewitness memory principles (memory loss, memory construction, misinformation effect, social influence, and confidence

eyewitness identification, and the legal system (*cont.*)
inflation), 145–147; and normative social influence, 146. *See also* eyewitness science; lineup reform
eyewitness science: and double-blind procedures, 157–158n6; and field identification, 155; and field studies, 154; and method, 154–156; and public policy, 153–154; and untidiness in policy development, 156–157
Eyewitness Evidence: A Guide for Law Enforcement (National Institute of Justice), 148
Eyewitness Evidence: A Trainer's Manual for Law Enforcement (National Institute of Justice), 148
Eyewitness ID Reform Act (2007), 151

Fagerlin, A., 357
Fair Labor Standards Act (FLSA [1938]), 271; and the distinction between an employee and an independent contractor, 272; and the failure to cover independent contractors, 272–273; minimum wage requirements of, 272
false convictions, 163, 176–177; and African Americans, 167; background of, 163–165; cases of (big and small), 175–176; causes and predictors of, 165–166; concerning rape and murder, 164; and DNA evidence, 164, 167, 168, 174; and the effects of appellate review, 177n7; and exoneration, 170, 177nn2–3; and exoneration as an official act, 164; and extraordinary relief, 170; and eyewitness error, 166; and the federal criminal justice system, 178n9; and forensic error, 166; frequency of, 165; and government misconduct, 175; and guilty pleas, 164–165; and ineffective defense work, 166; and light sentences, 164; and lineup identification, 174; and minority juveniles, 167; and perjury by informants, 166; and plea bargaining, 175–176; policy implications of, 173–176; and the production of evidence, 173–175; proposed solutions for, 177; and prosecutorial misconduct, 178n10; rate of, 177n4; and the recording of interrogations, 174; social and institutional context of, 167–168; and system variables, 166; and teenagers, 167. *See also* criminal investigation and adjudication, structure of; false convictions, and the adversary system
false convictions, and the adversary system, 170, 177–178n8; and confirmation bias, 170–171; and false claims of innocence, 171–172; and generating false negatives, 172–173; in Great Britain, 177–178n8; and the *juge d'instruction*, 170, 171; preparation for adversarial trials, 172
false hope syndrome, 368
Farmer, John, 151
Farmer, Paul, 138
"fast friend" paradigm, 38
Fazio, R. H., 37
Fechner, Gustav, 129
Federal Emergency Management Agency (FEMA), 411n1, 411n3
Federal Glass Ceiling Commission, 64
Federal Reserve Board, 448, 453
Federal Trade Commission (FTC), 421
Fehr, E., 270, 271, 272, 278n8, 498n2
Feigenson, N., 232
Fetherstonhaugh, D., 129–130
Fidelity Investments, 247
Finkelstein, S. R., 421
First Account program, 289
Fischbacher, U., 498n2
Fischbeck, P. S., 226
Fischhoff, B., 219, 221, 222, 226, 382
Fiske, S. T., 275, 276
floods/flooding, 86, 232, 382, 384, 386, 401, 403, 404, 411, 489
Flyvbjerg, B., 485
Foddy, M., 86
forced choice, 423, 431
forcing function, 432
formal procedures, characteristics of, 483
Fowler, J. H., 98
framing, 387; attribute framing, 384; in weight-loss programs, 370
Frankfurt School, 302
Frantz, C. P., 211
Freakonomics (Dubner and Levitt), 475
Frederick, S., 127, 181, 266
Freud, Sigmund, 15
Friedrich, J., 130
Friends Don't Let Friends Drive Drunk campaign, 303–304
Fryer, R. G., 330
fundamental attribution error, 199, 283
fungibility assumption, 283
Furman v. Georgia (1972), 163

Gäbel, H., 418
Gaertner, S. I., 35, 37, 43
Gaeth, G. I., 384
Gale, W. G., 375n3
game-theory, 119–120; *Bursa* game, 120; *Community* (*Kommuna*) game, 120; Prisoner's Dilemma game, 120; *Wall Street* game, 120
Garland, David, 187
Garret, B., 168
Gay, S., 418–419
gays, 22
gender bias, 17; ambivalent biases, 58; modern/subtle gender bias, 52–54; organization initiatives for solutions to, 61–67; policy tools for subtle gender bias, 60–61; subtle gender bias, 58. *See also* gender bias, subtle, controlling of
gender bias, subtle, controlling of, 58–59; failures of motivated control, 59; and the inefficiency of individual treatment adjudication, 60; summary of, 60; and the targets' inability to recognize and claim discrimination, 59–60
gender stereotypes, 15
General Accounting Office (GAO), 222
genocide, 133
Gerber, A. S., 92, 94, 95–96, 98, 102, 103, 151
Gestalt school of thought, 302
Geter, Lenel, 170
get-out-the-vote (GOTV), 91–92, 95, 96–97, 102–103, 103n1, 103n5; GOTV content and self-monitoring, 102; and identity-labeling, 100–101. *See also* get-out-the-vote (GOTV), modes of
get-out-the-vote (GOTV), modes of, 92; direct mail, 97; and the "foot-in-the-door" technique, 100; impersonal one-way communications, 93; interpretation of personal communications, 93–94; personal face-to-face contact, 92–93; personal phone calls, 93, 99–100; summary of, 94; and the use of digital technology, 94
Gigerenzer, G., 491n4
Gilbert, D., 355
Gilkeson, J. H., 253
Gimbel, R. W., 418
Gladwell, M., 406, 475
Glover, Jonathan, 126
Gneezy, U., 306
Goffman, E., 304
Goldstein, D., 261n10, 393, 417–418, 496, 498n5
Gollwitzer, P. M., 370
Good, C., 335
Goodnough, A., 399
Gordon, R., 458n19

Gourville, J. T., 374
Grassley, Charles, 207
Greathouse, S. M., 149
Gregg v. Georgia (1976), 163
Green, A. R., 41
Green, D. P., 92, 93, 94, 192
Green, K., 240–241n3
Greenberg, A., 372
Greene, J. D., 134
Greene, N., 185
Greenwald, A. G., 95
Gross, S. R., 165, 167, 169, 177n4
group polarization, 490
group-think, 111
groupthink (Janis), 490
Gruber, J., 276, 375n3
Gurnik v. Lee (1992), 267
Guthrie, C., 481

Haidt, J., 127, 182
Haisley, Emily, 367
Hamermesh, D. S., 361
Hamilton, D. L., 132
Hand, Learned, 163, 176
Hansen, J., 382–383, 383
Hardin, C. D., 275
Hardin, G. 381
Hardisty, D. J., 384
Hargreaves, D. J., 389
Harrington, Terry, 177n3
Harris, V. A., 199
Hasel, I. F., 147
Hastorf, A., 112
Hauser, M., 182
hazard prevention, psychology of, 400; biases in temporal planning (underweighing the future, underestimating risk, affective forecasting errors), 401–404; budgeting heuristics, 400–401; learning failures, 404–405; the politician's dilemma, 407–408; the Samaritan's dilemma, 406–407; social norms and interdependencies, 405–406
Heal, G., 406
health care: health-care decision aid (DA) developers, 352; and the inadequacy of decision making, 352–353; measurement of health-related utilities, 354; structure and evaluation of health care decision aids (DAs), 352
health-care policy, behavioral decision science application to (précis of book), 475; comprehension, 475–476; evaluation, 478–479; expression, 479–480; framework, 475; recall, 476–478
Heath, C., 198, 385
Heaton, P., 154
Hebl, M., 43
Hernandez, Alejandro, 170–171
Hersch, J., 382
Hertwig, R., 386
heuristics, 24n1, 261n2, 274, 400, 421; the availability heuristic, 232–236, 240, 383, 486; heuristic processes, 181; social amplification of, 240
Hewitt Associates, 247, 252
Higgins, E. T., 284
Hobbes, Thomas, 114
Holden, S., 257
Home Ownership and Equity Protection Act (HOEPA [1974]), 452
Hopkins, Nancy, 202
Howland, Carl, 302
Hsee, C., 233
Hsu, M., 385
Huang, M., 257
Huber, J., 403
Huber, O., 403
Huber, O. W., 403
Huberman, G., 261n2
Huff, C. R., 177n4
Human Genome Project, 475

identifiable victim effect, 373–374
identification: building of, 87–88; emotional identification, 80; social identification, 80
identity, 65, 79, 80, 81, 100, 101, 284, 303, 333, 334, 406; identity engagement process, 333-338; identity threat, 7, 333–334, 335, 338, 339, 341, 478; group identity, 23, 37, 39, 44, 61, 100, 111, 388; in-group identity, 37, 44, 61, 111; national identity, 388; self-identity, 303, 477; social identity, 80, 92, 98, 99, 100, 332, 333, 335, 338, 341. *See also* decision-making, and identity salience; identity engagement process; naive realism, and group identity experiments; poverty, and identity; social motivations, and identity; voting, as an expression of one's identity/self-expression
identity engagement process, 333–338; educational intervention, 335; removing social identity at the vigilance stage, 335–336
Illinois Eyewitness Identification Field Study, 153
implementation intentions, 370
Implicit Association Test (IAT), 55–56, 274
implicit prejudice, 13–14, 23–24; characteristics of, 15–17; discovery of, 14–15; nature of, 14; predictive validity of, 17. *See also* implicit prejudice, consequences of implicit prejudice, consequences of, 17; effects of social dynamics on, 21; implicit effects of cognitively accessible stereotypes/prejudices, 17–18; implicit prejudice as cognitive associations, 18–20; implicit prejudice and the medical profession, 19–20; implicit prejudice as subject to interpersonal dynamics and social influence, 21–23; implicit prejudice and voting behavior, 19; and the predictive validity of implicit attitudes, 19; social control of implicit prejudice, 20–23
incentives, 63, 78, 79, 85, 86, 198, 218, 264, 308, 369–370, 437, 440; economic/financial incentives, 198, 304, 307, 364, 372, 409, 411, 425, 444, 445, 458n4, 478; employer incentives, 458n3; incentives for firms, 44, 445, 458n4; market incentives, 445, 446, 450, 451, 457n1; material incentives, 85, 308, 391; and prices, 437–438; psychological incentives, 301, 308; tax incentives, 410, 440, 446, 472. *See also* choice architecture
independent contractors, 272–273
individual development accounts (IDAs), 368
Individual Retirement Accounts (IRAs), 265
information cascades, 233, 240, 406, 485
informed consent, 223
Informedix Med-eMonitor System, 371
Innocence Inquiry Commission, 177–178n8
Innocence Project, 145, 151, 157n3
Institute of Medicine, 41, 45; report of (2003), 32–33
institutions, 442–443; institutions provide implicit planning, 443–444; institutions shape behavior, 443; institutions shape defaults, 442–443. *See also* institutions, and financial access
institutions, and financial access, 285; institutions provide implicit planning, 285–286; institutions shape behavior, 285; institutions shape defaults, 285; popularity of institutions among the poor, 287–288
instrumental models, 79
instrumental motivations, 83; dependence, 83; distributive fairness, 83; environmental contingencies, 83; instrumental trust, 83; investment, 83

intergroup contact, 45
intergroup dialogue. *See* conflict resolution, methods for employing
intergroup dynamics, 45–46
intervention(s), 301–302, 303, 307; educational intervention, 335, 336–338; and social-norms marketing, 304
intransigence, 108
Inzlicht, M., 335
Israel, day-care centers in, 306
IXI Company, 262n16
Iyengar, S., 245–246, 441

Jackson, H. E., 451
Janis, Irving, 490
Janoff-Bulman, R., 211
Jepson, C., 355, 357
Jiang, W., 261n2
Johnson, E. J., 261n10, 384, 385, 393, 417–418, 431, 468, 479, 496, 498n5
Jolls, C., 272, 275, 278n5
Jones, E. E., 199
judgmental errors, 481, 486
Just, R., 138
Just and Painful: A Case for the Corporal Punishment of Criminals (Newman), 187
justice: distributive justice, 80; intuitive origins of, 182; procedural justice, 80, 88n1
Justice Project, 151

Kahn, B. E., 312
Kahneman, Daniel, 78, 117, 127, 129, 181, 251, 267, 270, 276, 363, 475, 489
Kaley, A., 64
Kam, C. D., 98
Kamenica, E., 245–246
Kamlani, K. S., 270
Kansas v. March (2006), 165
Kant, Immanuel, 183
Kassin, Saul, 166
Kawakami, K., 35
Keeney, R. L., 240–241n3
Kelly, F., 64
Kelman, H. C., 80, 83, 209
Kennedy, K. A., 200, 204, 210
Kernochan, J., 182
Kerry, John, 98
Kimmel, S. E., 371
King, E. B., 43
Kirchsteiger, G., 270, 272, 278n8
Klein, W., 232
Kling, J. R., 291, 490
Klobuchar, Amy, 151, 154
Knetsch, J. L., 251, 267, 270, 276
Koehler, D. J., 441
Kogut, T., 132
Kohlberg, H., 182
Korobkin, R., 481–482
Kotlikoff, L. J., 253, 255
Kovera, M. B., 149
Krieger, L. H., 275, 276
Krishna, A., 421
Krishnamurti, T. P., 221, 222
Kristof, Nicholas, 131
Kugler, M. B., 200, 210, 211
Kunreuther, H., 386–387, 399, 400, 406, 407, 408, 411n4
Kydland, F., 408

Lanham Act (1946), 458n8
Larimer, C. W., 102
Laska, S. B., 399, 411n1
Latino Americans, 329, 331
law: age-discrimination law, 264, 269; social demand for, 236. *See also* employment law
LeBlanc, A. N., 289
legislative fact-finding, 485–486; and the affect heuristic, 486; and availability and related biases, 486; and biases in social perception stereotyping, 487; and distortions of facts by advocates, 487; and leakage of trade-off preferences into fact-finding, 487–488; and overconfidence, motivated skepticism, and confirmation bias, 487; and poor grasp of probability statistics and empirical methodology, 486; and psychic numbing, 486–487
legitimacy, 80
Lein, L., 286
Leo, R., 169
Lepper, M. R., 109–110, 112, 210, 441
Lerner, J., 403, 488
lesbians, 22
Lessig, Lawrence, 305
levee effect, 403
Levin, I. P., 384
Levine, D. I., 270
Levitt, Steven, 475
Lewin, Kurt, 23, 88, 302, 330
Liaison Committee on Medical Education, 41
Liberman, N., 189, 384, 404
Liberman, V., 121
Lichtenberg, J., 498n7
Lichtenstein, S., 382
Liesch, M. J., 421
lifestyle diseases, 361
Lifton, R. J., 133
Lin, D. Y., 198, 204
Lindemann, P. G., 389–390
Lindsay, R. C. L., 156
lineup reform, 150–152, 156, 157nn1–2; and mock witness procedure, 157n5; resistance to, 152–153; in Suffolk County (Boston), 151. *See also* lineup reform, key events in the growth of; lineup reform, protocol for
lineup reform, key events in the growth of: DNA exonerations, 148; and the legal environment, 147; and the National Institute of Justice guide, 148; system and estimator variables of, 148
lineup reform, protocol for: double-blind sequential lineups, 149–150; relative and absolute judgment, 148–149
List, J. A., 253
Little Rock Central High School, 13
Living High and Letting Die (Unger), 373
Loewenstein, George, 131, 132, 202, 206, 266, 355, 367, 404
Loomes, G., 387
Lord, C. J., 109–110, 210
loss aversion, 115–116, 117, 235, 283, 291, 363, 369, 383–384, 387, 421–422, 489; nominal loss aversion, 250–252
lottery-linked savings accounts, 367–368
Lovallo, D., 363
Lowery, B. S., 22, 275
Lusardi, A., 253
Lutter, R., 240–241n3

MacArthur Justice Center, 153
"mad cow disease," 233
Madrian, B. C., 245, 246, 366–367
Magat, W., 403
magnetic resonance imaging (MRI), 18; functional magnetic resonance imaging (fMRI), 62
Majumdar, B., 41
mandated choice. *See* forced choice
Mankiw, N., 255
Mann, R., 453
Mann-Whitney Wilcoxon rank test, 258
Maoz, I., 117
mapping, 434–435
marginal propensities to consume (MPC), 283
Marx, Karl, 22
Marx, S., 382–383, 383
Massachusetts Institute of Technology (MIT), 202
Mathews, J., 343
McClelland, G., 403
McClesky v. Kemp (1987), 486
McGhee, Curtis, 177n3

McKendrick, J., 389
McKenzie, C. R. M., 421
McNeil, B. J., 353
Mecklenburg Report, 152, 153
Medicaid, 39
medical care: continuity of, 41; patient-provider "teams," 42
Medicare, 39; Medicare Part D prescription-drug coverage, 291, 471
medicine: cost of, 173; medical care and the use of checklists, 173; practice of, 19–20
Mendoza-Denton, R., 38
mental-models protocol, 222
mentoring, 43
Merton, R. C., 257
Merz, J., 223
meta-analysis, 154; estimator variable meta-analysis, 157n4; peer-reviewed meta-analysis, 154
Meyer, R., 405
Michel-Kerjan, E., 411n4
Middleton, J. A., 93
Milch, K. F., 388
Milgram, Stanley, 2, 282, 440–441
Miller, D. T., 198
Mindless Eating Challenge, 319–321; method of, 319–320; results of, 320–321
Mindless Eating: Why We Eat More Than We Think (Wansink), 319, 325
minimum wage law, costs of, 273
misfearing. *See* cost-benefit analysis, and correction of misfearing
Mitchell, O. S., 253
Mnookin, L., 115
Moldonado-Molina, M. M., 188
Molina, Otto Perez, 189–190
Moore, D. A., 201, 206
moral intuition, 127; failure of, 133–134
moral judgment, 127
Morgenstern, O., 354
motivated skepticism, 487
motivation, in organizations, 61–64; distinction between cognition and motivation, 77–78; effectance motive, 63; harnessing the motive to belong, 61–62; harnessing the motive to control, 62–63; harnessing the motive for self-enhancement, 63; harnessing the motive for trust, 63–64; harnessing the motive to understand, 62; instrumental models for, 79; and judgments, 77. *See also* social motivations
motivational dynamics, 301–302; approach motivation, 302; avoidance motivation, 302
MoveOn.org, 93
Move to Opportunity program, 330–331
Mullainathan, S., 447, 453, 490
multi-attribute utility theory (MAUT), 488
myopia, 287, 288, 290, 381, 383, 393, 399, 408; cognitive myopia, 383, 489; planning myopia, 402; temporal myopia, 456

Nadler, J., 185
Nagel, T., 498n6
naive realism, 108, 200, 208, 364; and apparent versus real differences in basic values, 113–115; and biased perception and perceptions of bias, 111–112; convictions of the naive realist, 110–111; and the false consensus effect, 110; and false polarization, 112–113; and group identity experiments, 111; and the hostile media/mediator effect, 112; and perceptions of bias, 108–110. *See also* naive realism, lessons from the real world
naive realism, lessons from the real world: conversion from militant to peacemaker, 123; futility of convincing people of something they cannot "afford" to understand, 123; importance of relationships and trust, 121–122; importance of a shared view, 122
National Academy of Sciences, 82
National Association of Criminal Defense Lawyers, 153
National Flood Insurance Program (NFIP), 409
National Restaurant Association, 325
National Training Laboratory, 302
natural disasters, 398–399, 411; Alaska earthquake (1964), 405, 407; Cyclone Nargis, 398; Hurricane Andrew, 407, 408; Hurricane Ike, 398; Hurricane Katrina, 398, 402, 407; Hurricane Wilma, 404, 405; hurricanes in Florida, 399, 408; and the investment-mitigation gap, 399–400; Southeast Asia tsunami (2004), 398; Tropical Storm Agnes, 407. *See also* hazard prevention, psychology of; strategies for building resilient communities prone to natural disasters
negotiation, 208; and negotiation impasse, 364
New Deal, 109
Newman, Graeme, 187
NIMBY (not-in-my-backyard) phenomenon, 117, 306–307
Nisbett, R., 183, 283
No Child Left Behind Act (2001), 419–420, 431
Norman, Donald, 429, 432
normative theory, 401, 403
North, A. C., 389
North Carolina Actual Innocence Project, 151
Northern Ireland, 203–204
nuclear power, 232, 234, 235, 238, 241n4, 245, 381, 488
Nudge (Thaler and Sunstein), 467, 490
nudges, 467, 468–469
nullification, and jury trials, 185

Obama, Barack, 457
objectivity, 211; problems associated with, 211. *See also* objectivity, strategies for implementing
objectivity, strategies for implementing, 211; demand objective behavior, 212; educate about unconscious bias, 211–212; reduce exposure to biasing information, 212
obligation, 80
O'Brien, B., 165, 167, 171, 175, 177n4
O'Donoghue, T., 266, 277n2, 404
Office of the Comptroller of the Currency, 453
Office of Federal Contract Compliance Programs (OFCCP), 64–65
office supplies, personal use of, 307–308
ordinal equity, 59
organ donations, 261n10, 417–418, 431
organizations: organizational design, 84, 84–87; organizational initiatives, 64. *See also* monitoring/motivation, in organizations
Orlich, A., 95–96
Orszag, P. R., 375n3
Orwell, George, 109
overoptimism, 365
over-the-counter (OTC) sales, 221, 222
Ozanne, J. L., 284

Page-Gould, E., 38
Palm, R., 399
Pareto inferiority, 277n2
Parnes, Lydia, 421
Patenaude, Pat, 150
paternalism, 282; asymmetric paternalism, 362, 467–468; hard paternalism, 495; and manipulation, 494–496; soft paternalism, 495
Patnoe, S., 340
Patraeus, David, 308
Pauly, M., 408

pay-as-you-drive-and-you-save (PAYDAYS) insurance program, 372–373
peanuts effects, 264, 375n2
Pearson v. Shalala (1999), 219
Pennebaker, J., 336
Penner, L. A., 41
Pension Protection Act (PPA [2006]), 246, 249, 252, 465
People v. Collins (1968), 486
perspective taking, 208; problems associated with, 208–209. *See also* perspective taking, strategies for implementing
perspective taking, strategies for implementing, 209; "consider the opposite" strategy, 210; encouraging cooperative norms, 209; limiting counterarguments, 210; manipulating visual perspective, 209; using carefully worded instructions, 209–210
Peters, E., 132, 382
Pew Research Center (2009 opinion poll, climate change), 382
pharmaceutical industry, relationships with physicians, 195–196, 201–202
physical exercise, 187, 247–248, 268, 301, 326, 370–371, 429, 438, 479
physicians: and the materiality standard when securing informed consent, 223; physician-patient relationships, 41. *See also* decision aids (DAs); health care; pharmaceutical industry, relationships with physicians
Piaget, Jean, 314
Pickering, A., 154
Pitinski, T. L., 284
planning myopia, 402
Plaut, V. C., 46
Plouffe, David, 430
Podesta, G., 387
Policy and Choice (Congdon, Kling, and Mullainathan), 490
policy makers, 482
politics, power, and freedom, 497
Polivy, Janet, 368
Poon, C. S. K., 441
postcompletion error, 432
poverty, 281–282, 293–294; alleviation of, 473–474. *See also* poverty, behavioral perspectives on; poverty, and economic behavior of the poor; poverty, and institutional access for the poor; poverty, noninstitutional aspects of
poverty, behavioral perspectives on, 291–292; and channel factors, 283; and cognitive load, 292; and construal, 282; and context dependence, 282; and identity, 283–284; and limited mental resources, 292; load, stress, and tunneling aspects of, 292–293; and mental accounting, 282–283
poverty, and economic behavior of the poor: behavioral perspective of, 289; and check cashing, 291; and mental accounting schemes, 289–290; and the "unbanked," 288–289; and the use of payday loans, 290–291
poverty, and institutional access for the poor: features of financial access, 285–286; role of financial access, 284–285; and rotating savings and credit associations (ROSCAs), 288
poverty, noninstitutional aspects of, 286; and high-interest-rate borrowing, 286–287; and lack of financial slack, 286–287; and layaway plans, 288; and no-buffer stock savings, 288; and SEED commitment savings, 288; and small to large transformations, 287–288
precautionary principle, 388
preference reversal, 266
preferences, 420
preferences-as-memory framework, 385
prejudice, explicit, 16
Prescott, F., 408
presumed consent, 418, 431
Previtero, Alessandro, 252
probability neglect, 486
productivity, 77, 269, 319, 323
Pronin, Emily, 111–112, 198, 200, 204, 210, 211
prospect theory, 117, 129, 364, 372, 383–384, 385, 386, 387, 489
psychic numbing, and mass atrocity, 126–127, 129–130, 139, 477; and the collapse of compassion, 131–133; and the failure of moral intuition, 133–134; implications for international law and policy, 134; numbers and numbing, 130–131; psychological lessons concerning, 127–128. *See also* psychic numbing, strategies for dealing with
psychic numbing, strategies for dealing with, 134; changing the method and content of human rights reporting, 136; constructing default rules and precommitment devices, 134–135; direct promotion of System 2 deliberation, 138–139; early warning and prevention action, 135; employing System 1 to support System 2 processes (use of affective imagery), 137–138; empowerment of institutions, 135–136; reconsidering elements of human rights law, 137; reconsidering human rights indicators, 136–137; and victim empowerment, 138
psychological environment, 294, 331, 332–333, 342
psychology, 1; interface of with economics, 78; of interracial interactions, 33; role of in public policy, 1. *See also* cognitive psychology; social psychology
"psychometric paradigm," 237
public health, 221, 307, 325, 375, 479; improvements to, 325–326; public health officials, 225, 226
public policy, 1, 5, 6, 9, 13, 15, 20, 21, 23, 77–78, 81, 84, 108, 109, 145, 156, 165, 231, 238, 465, 470; and economics, 361, 365, 424; effectiveness of, 46; role of human behavior in, 1; and science, 153–154; success in public policy debates, 109. *See also* competence, and public policy
punishment judgments, 181; and developing disrespect for the law, 194–185; and the gap between legal codes and community settlements, 185; just-desert intuition and the policy consequences for legal codes, 184; and the policy-capturing approach, 183; punishment decisions as intuitive, 181–182; punishment intuitions as just-deserts based, 183–184; and the trolley-car scenario, 182
punishment/retribution/deterrence, behavioral issues of, 181; general deterrence, 183; high and low need for deterrence, 184; specific deterrence, 183. *See also* deterrence, by increasing sentence duration
Putnam, Robert, 13

Qualified Residential Mortgages, 452
query theory, 385, 392, 393
Quinn, D. M., 332

Rabin, M., 277n2, 404
race, 17; and the recategorization of racial identities, 39; university residential housing and race relations, 35–38. *See also* racial bias; racism
racial bias, 23, 32–33; intergroup bias, 33–34; and intergroup contact, 34–35; subtlety of, 40, 45. *See also* discrimination, racial; racial bias, in context

racial bias, in context, 35; in an educational context (university residential housing), 35–39; in a medical context, 39–42; in a workplace context, 42–45
racism: aversive racism, 33–34; persistence of, 202–203; and "sensitivity training," 22; and the White Castle killing incident, 195
Radelet, M. L., 163
rape, 57, 163, 164, 165, 166–167, 169, 238; role of DNA in rape convictions, 164, 174
rational agent model, 4, 440, 444
rationality, 133, 375; bounded rationality, 6, 274, 265, 274, 277, 380, 386, 478; economic rationality, 400; and the good, 494–495; and morality, 497; procedural rationality, 482; "richer rationality," 237, 238; "rival rationality," 236, 237; unbounded rationality, 389
Ratner, R. K., 198
Read, D., 247
Real Estate Settlement Procedures Act (RESPA [1974]), 452
recall, 476–478
RECAP (Record, Evaluate, and Compare Alternative Prices), 435
recursion, 335; negative recursion, 335; recursive performance cycles, 337, 338
Redelmeier, D., 441
Reeder, G. O., 200
regret aversion, 368, 369, 372
regret theory, 387
regulation, 81, 220, 232, 233, 235–236, 282, 284; behaviorally-informed regulation, 8, 9, 440, 446, 457; legal regulation, 265, 271, 273; minimum wage regulation, 6, 265, 270, 273; pension regulation, 6, 265, 277, 446, 458n3; product regulation, 444, 446, 449, 450; systemic effects of, 234–235; under the Reagan administration, 241n8. *See also* behaviorally informed regulation; regulatory capture
regulatory capture, 362
Reilly, Peter, 170
reinforcement learning, 404–405
Reno, Janet, 148
required choice. *See* forced choice
Research Center for Group Dynamics, 88
retirement security, 472
Richeson, J. A., 19, 36, 275
Riedl, A., 250, 272, 278n8
"rights entrepreneurs," 66
Rips, L. J., 388
Risinger, D. M., 165
risk, 86, 111, 117, 129, 130, 133, 233, 238, 362, 398–399, 486, 489; of bias, 212; at-risk children, 329, 343; dread risk, 382; emotional response to risk, 234; judgments concerning risk severity, 238; relevant risk, 236, 237; risk assessment, 128, 133, 467, 471; risk aversion, 363, 365, 383, 491n5; risk-based funds, 246, 252, 253, 255, 258, 260–261; risk and benefits, 235; risk reduction, 399, 409; risk regulation, 235, 240; risk taking, 8, 245, 246, 257, 258, 260, 386, 398, 491n5; unknown risk, 382
Ritov, R., 132
Robinson, P. H., 183, 184, 188
Robinson, R., 113
Rogers, T., 98, 102, 103
Roosevelt, Franklin D., 109, 126
Roper v. Simmons (2005), 217
Ross, A., 117, 204
Ross, L., 109–110, 112, 115, 117, 121, 198, 210, 283, 364
rotating savings and credit associations (ROSCAs), 288
Rothengatter, W., 485
Rottenstreich, Y., 233
Rudman, L. A., 18, 22
Rustichini, A., 306
Rutherford, S., 287
Ryan, George, 152

SAFE Act (2008), 452
Safelite Group, 249–250
Salvatore, J., 35
Samuelson, P. A., 249
Samuelson, W. E., 257
Sanchez-Burks, J., 44
Sarbanes-Oxley Act (2002), 489
SARS, 232
Save the Children, 131, 132
Save More Tomorrow (SMarT), 247, 248, 249, 261n5, 268, 367, 374, 419
Scalia, Anthony, 165, 207
Schacter, Daniel, 146, 155
Schelling, Thomas, 373, 406
Schleh, Jeanne, 151
Schliemann, A. D., 341
Schlup v. Dello (1995), 170
School Nutrition Association, 325
Schroeder, S. A., 368
Schulman, K. A., 39
Schulze, W., 403
Schwartz, L. M., 221
Securities and Exchange Commission (SEC), 453
self-affirmation, 336, 338–339
self-interest, 61, 77–79, 86, 111, 112, 114, 195, 196, 198, 201, 202, 206, 209, 340, 374, 478, 494; bounded self-interest, 6, 264, 270, 277, 478; financial self-interest, 195, 198, 201, 206; material self-interest, 78, 79, 264, 270; self-interest biases, 198
self-judgment, and behavior, 18
sensitivity training, 41; for sexism, 22
sequential superiority effect, 156
serial semantic priming paradigms, 16
sexism, 22, 53, 195; ambivalent sexism, 58; benevolent sexism, 58; hostile sexism, 58; persistence of, 202–203; subjectively benevolent sexism, 58. *See also* gender bias
Shafir, E., 441, 447, 453, 488
Shafir, S., 386
Shang, J., 374
Shaw, George Bernard, 114
Shea, D. F., 245, 246, 366–367
Shefrin, H. M., 266
Shelton, J. N., 19, 35, 36
Sherif, Muzafer, 23
Sherman, S. J., 95, 132
Shif, M., 284
Shook, N. J., 37
Siegelman, Peter, 60
Silberman, M., 185
Simmons v. United States (1968), 147
Sinclair, S., 275
Singer, Peter, 133
Skagerberg, E. M., 154
Skinner, B. F., 22, 23
Slater, A., 154
Slichter, Sumner, 270
Sloate, P. L., 390
Slovic, P., 89, 130, 131, 132, 382, 386–387
Slovic, S., 130
Small, D. A., 131, 132
Small Plate Movement, 325
Smarter Lunchrooms Project, 325
smart-grid technology, 392
Smith, Adam, 375
Smith, J. K., 95–96
Sobel, J., 316
Sober, E., 498n2
social categorization, 33, 38
social cognition, dual process model of, 62
social identification, 80
social influences, 489–490; escalation of commitment, 490; group dynamics, 490; reciprocity, 490; social proof (conformity), 490
social judgment, 14
social motivations, 79; analysis of, 83–84; attitudes, 79, 83; distinction of from instrumental motivations, 86;

social motivations (*cont.*)
and identity 80, 83; and motive-based trust, 80, 83; policy implications of, 84; and procedural justice, 80, 83; and social values, 79–80, 83; weaknesses of, 86–87; why the strategy of motivation matters, 81. *See also* instrumental motivations; social motivations, impetus of; social motivation strategies, empirical comparisons of social motivation strategies, empirical comparisons of: analysis of, 82; cooperation, 82; cooperation in work settings, 83; design, 82; policy implications of, 82–83; regulatory settings, 81–82; social motivation, 82
social motivations, impetus of, 80–81; content of the concerns people express, 80; and the influence of social motivations on cooperative behavior, 81; social motivation as empirically distinct from indicators of material gain, 80–81; social motivation and the production of consistent behavior, 81
social organization, 20; changes in, 21
social policy, definition of, 1
social psychology, 6, 14, 21, 23, 84, 98, 281, 301, 338; cognitive social psychology, 62; contributions to the understanding of disagreement, 108–109
social scientists, 302
social security system, 109, 264, 268–269; and employer-employee agreement, 269
social tension systems, 329–331
social values, 79–80
Society for Human Resource Management, 64
Solow, Robert, 270
Son Hing, L. S., 43
Sontag, Susan, 138
Soss, N. M., 361
South African Constitutional Court, 139
Spencer, Steve, 332
Sprinzak, Ehud, 203
stakeholders, 87
Stalin, Joseph, 486–487
Stand and Deliver (Mathews), 343
Stanford Center on International Conflict and Negotiation (SCICN), 108, 111
Stanovich, K. E., 127
startle response, 16
Steblay, N., 154
Steele, Claude, 332
Steinberg, T., 408
stereotypes/stereotyping, 23–24, 33, 62; stereotype activation, 62–63; stereotype salience, 15; "stereotype threat," 284; stereotypes as double-edged swords, 18; ubiquity of stereotyping, 15. *See also* gender stereotypes; stereotyping, ambiguous; stereotyping, automatic
stereotyping, ambiguous, 57–58; summary of, 58
stereotyping, automatic, 54; and accessibility (priming effects), 54–55; and category activation, 56–57; and category confusion, 54; and the Implicit Association Test (IAT), 55–56; summary of, 57
Stevens, S. S., 129
sticky opt-out mortgage system, 449–451
Stillinger, C., 115
Stovall v. Denno (1967), 147
strategies for building resilient communities prone to natural disasters, 408; long-term insurance (LTI) and long-term mitigation loans, 409–410; seals of approval, 410; tax incentives, 410; zoning ordinances that better communicate risk, 410–411
Stroop, J. R., 439
Strotz, R. H., 247
subjective expected utility (SEU), 481
Sugden, R., 387
Sullivan, Thomas, 153
Sunstein, C. R., 245, 362, 421, 438, 467, 490, 495, 498n5
Susskind, J., 132
Susstein, Cass, 66, 305
System 1-System 2 thinking, 127, 128, 133, 231, 274, 137–139, 381, 438n2, 483, 489
Szent-Györgyi, Albert, 129

T. Rowe Price, 247, 249
taxes, 304–306; calibrating taxes and subsidies, 308
terrorism, 204, 232
Tetlock, Philip, 485
T-group movement, 302
Thaler, Richard H., 245, 246, 251, 253, 257, 261n2, 264, 265–266, 267, 268, 270, 276, 362, 367, 374, 383, 393, 421, 429, 433, 438, 467, 490, 495, 498n5
third-person effect, investigations of, 112
threat: collective threat, 333; identity threat, 332–333; lessening the impact of social identity threat, 336–338
TIAA-CREF, 247
Titmuss, Richard, 306
Tirole, J., 375n1
Tollestrup,P. A., 154
"top-down" products of ideology, 112
TOPS (Take Off Pounds Sensibly) weight loss group, 325
Torelli, P., 330
Towles-Schwen, T., 37
Trail, T., 36
Training and Standards Bureau (Wisconsin Department of Justice), 151
transaction costs, 255, 276, 284, 420, 444, 471
transitivity, 482, 491n3
Trope, Y., 189, 384, 404
Tropp, L. R., 38
trust, 63–64, 88nn1–2; instrumental trust, 83; motive-based trust, 80, 83; "trust" relationships, 271
Truth in Lending Act (1968), 448, 498n4
Tufano, P., 453–454
tunneling, 292–293
Turtle, J. W., 154
Tuttle, William M., Jr., 185
Tversky, A., 78, 117, 129, 385, 436, 489
Tyler, T. R., 83, 84, 86, 88n1, 184
Tyndall Report, 138

Ubel, P. A., 355, 357
Uhlmann, E., 198, 211
Unger, Peter, 133–134, 373
Uniform Commercial Code, 448, 450
United Nations Security Commission, 135
United Nations Universal Declaration of Human Rights, 128
United States, 252; expansion of the prison population in, 186–187; number of felony convictions in, 176; regulation in, 81; and the risks of terrorism, 232
United States v. Garsson (1923), 163, 176
United States v. Wade (1967), 147
University of Massachusetts-Amherst, 38
U.S. Department of Agriculture (USDA), 324, 325
U.S. Department of Health and Human Services (HHS), 41
U.S. Department of Labor, 42, 252
U.S. Environmental Protection Agency (EPA), 240–241n3
U.S. Food and Drug Administration (FDA), 219, 221, 222–223

U.S. Small Business Administration, 407
utility analysis, 406, 476
utility theory: expected utility theory, 264, 265, 274, 276, 383–384, 385, 400, 401; multi-attribute utility theory (MAUT), 488

Vallone, R. P., 112
Valentine, T., 154
value of information (VOI), 387
van Ittersum, K., 314
Van Laar, C., 37
van Leeuwen, B., 247
Vanguard Investments, 247, 250, 253, 255, 256; Life Strategy Funds of, 253; OneStep Save program of, 248–249; Target Retirement Funds of, 253
Vasquez, David, 163
Västfjäll, D., 132
Vaughn, D., 168
Viceira, L. M., 257
Virginia Department of Forensic Science, 165
Virginia Tech University shootings (2007), 405
Viscusi, K. W., 382, 412
Vissing-Jorgensen, A., 255, 257
Volpp, K. G., 368–369, 371
von Hirsch, A., 187
von Neumann, J., 354
voting, 91–92; challenges to traditional modes of, 101–102; enhancing voter turnout, 99, 103n1; estimating net cost per vote, 103n3; voter eligibility, 94; voting behavior, 19, 91; voting as a static, self-interested decision, 94. *See also* get-out-the-vote (GOTV); voting, as a dynamic constellation of behaviors; voting, as an expression of one's identity/self-expression; voting, as a social act
voting, as a dynamic constellation of behaviors, 94; and implementation intentions, 96–97; and the "mere measurement effect," 95; and the nature of errors in prediction, 95; and pre-voting self-prediction and commitment, 94–96; social pressure and accountability after an election, 97; and the "self-prophecy effect," 95
voting, as an expression of one's identity/self-expression, 91, 99; and the "foot-in-the-door" technique, 99–100; and identity labeling, 100–101; and visual perspective, 101
voting, as a social act, 97–98; and social norms, 98–99; voting for the sake of others, 98
Vovat, G., 390
Vygotsky, L. S., 340

Wade-Benzoni, K. A., 388
Wagenaar, A. C., 188
Walton, G. M., 336
Wang, C., 357
Wansink, B., 312, 313, 314, 316
Ward, A., 117, 120, 364
Washington State park fees, 430–431
Wazana, A., 201
Weber, E. H., 129
Weber, E. U., 382–383, 383, 384, 385, 386, 389–390, 390, 390–391
Weber, Elke, 84
Weber's Law, 129
Webster, C. M., 187
Welch, H. C. G., 221
welfare reform, 199
Wells, G. L., 147, 151, 166
West, R. F., 127
West, T., 36, 37, 39
White, G. F., 386–387
White, Gilbert, 403
White Castle, killing of an African American by police at, 195
Wider, R., 403
Wikileaks, 137
Wilson, D. S., 498n2
Wilson, J. Q., 183
Wilson, T., 355
Wilson Center on the Hill, 487
Woloshin, S., 221
Wolz, C., 241–241n3
Woodhead, M., 339–340
Wooldridge, J. M., 255
Word, C. O., 43
words: "prime words," 15; "target words," 15
Wright, D. B., 154

Yale Communication and Attitude Change Program, 302
Yamaguchi, T., 261n15
Yellen, J. L., 270, 271
yield spread premiums (YSPs), 451, 452, 458n11
Young, L., 182
Yuille, J. C., 154

Zanna, M. P. 43
Zeckhauser, R. J., 249
Zeldes, S., 255, 257–258
Zelizer, B., 138
Zikmund-Fisher, B., 353
"zone of proximal development," 340