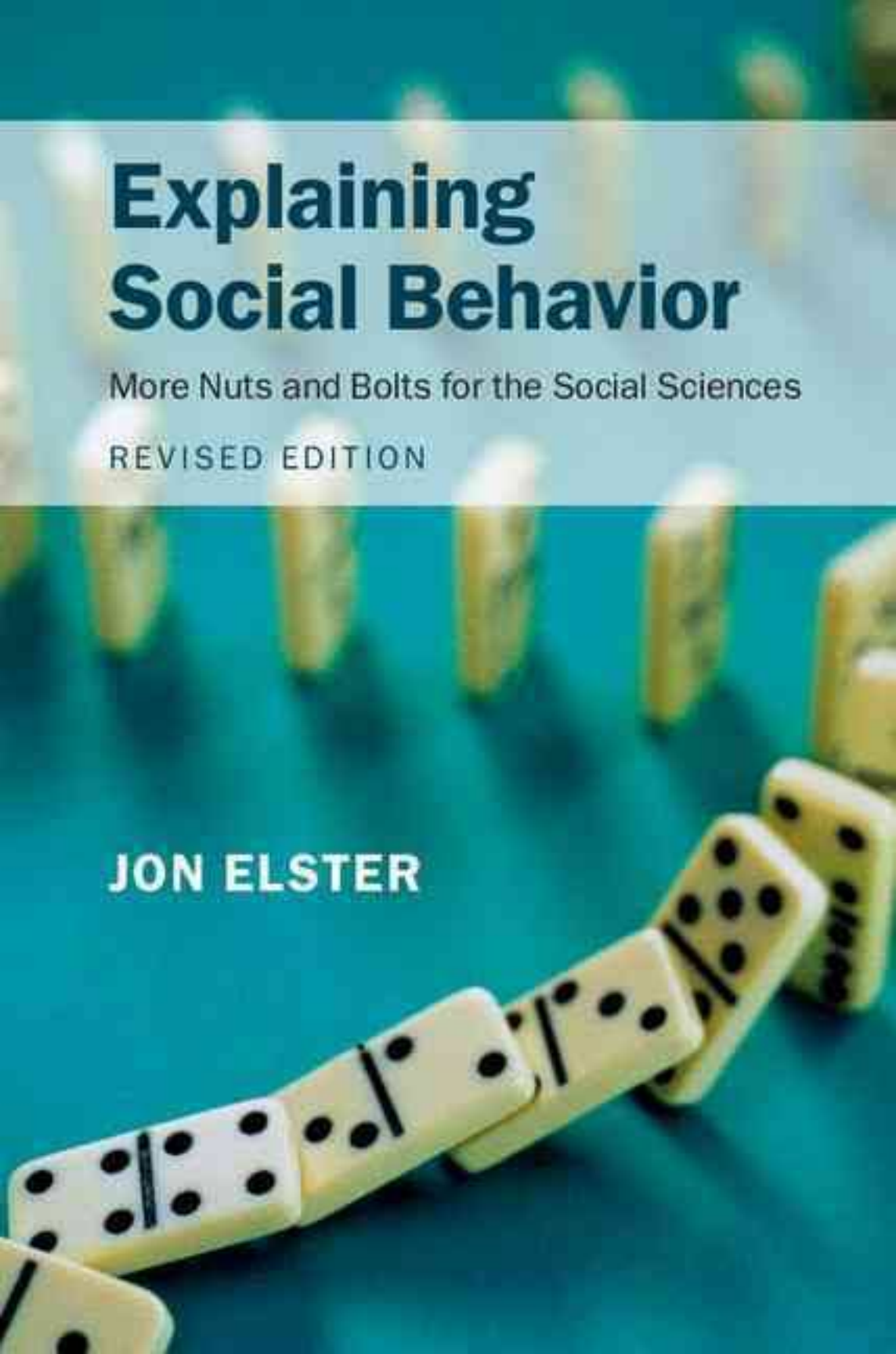
# Explaining Social Behavior

More Nuts and Bolts for the Social Sciences

REVISED EDITION



JON ELSTER

# Explaining Social Behavior

*More Nuts and Bolts for the Social Sciences*

Revised edition

In this new edition of his critically acclaimed book, Jon Elster examines the nature of social behavior, proposing choice as the central concept of the social sciences. Extensively revised throughout, the book offers an overview of key explanatory mechanisms, drawing on many case studies and experiments to explore the nature of explanation in the social sciences; an analysis of the mental states – beliefs, desires, and emotions – that are precursors to action; a systematic comparison of rational-choice models of behavior with alternative accounts, and a review of mechanisms of social interaction ranging from strategic behavior to collective decision making. A wholly new chapter includes an exploration of classical moralists and Proust in charting mental mechanisms operating "behind the back" of the agent, and a new conclusion points to the pitfalls and fallacies in current ways of doing social science, proposing guidelines for more modest and more robust procedures.

JON ELSTER is Robert K. Merton Professor of Social Science at Columbia University and Professeur Honoraire at the Collège de France. He is the author or editor of thirty-four books, most recently *Agir contre soi: la faiblesse de volonté* (2007), *Le désintéressement: traité critique de l'homme économique* (2009), *Alexis de Tocqueville: The First Social Scientist* (Cambridge, 2009), *L'irrationalité* (2010), and *Securities against Misrule: Juries, Assemblies, Elections* (Cambridge, 2013).

# Explaining Social Behavior

*More Nuts and Bolts for the Social Sciences*

*Revised edition*

Jon Elster
*Columbia University*

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

Cambridge University Press is part of the University of Cambridge.

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

www.cambridge.org
Information on this title: www.cambridge.org/9781107416413



First published 2015

Printed in the United Kingdom by Clays, St Ives plc

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

*Library of Congress Cataloging in Publication Data*
Elster, Jon, 1940–
Explaining social behavior : more nuts and bolts for the social sciences / Jon Elster. – Revised edition.
pages cm
Includes index.
ISBN 978-1-107-07118-6 (Hardback) – ISBN 978-1-107-41641-3 (Paperback)
1. Social sciences–Methodology. 2. Social interaction. I. Title.
H61.E434 2015
302–dc23 2015008862

ISBN 978-1-107-07118-6 Hardback
ISBN 978-1-107-41641-3 Paperback

To the memory of Aaron Swartz

# Contents

# Preface

The first edition of this work was an extension of a much shorter book, *Nuts and Bolts for the Social Sciences*. By and large, the extension was in breadth, not in depth. Many more topics were covered, but at more or less the same level of analysis. This revised edition covers roughly the same topics as the first, but provides, I hope, greater insight.

To make room for the substantial amount of new material, while keeping the book within a manageable size, Part IV – "Lessons from the natural sciences" – has been eliminated. Some discussions in that Part have been incorporated into the new Chapter 11, "Reinforcement and selection," and Chapter 20. A new chapter on "Transmutations" is added. The chapters in Part V on collective belief formation, collective action, collective decision making, institutions and constitutions, as well as the Conclusion, are entirely rewritten. Most chapters in Part II are also substantially modified. Parts I and III are also revised, but less heavily.

The revisions and additions draw on five books I have published in the meantime: *Agir contre soi* (2007), *Le désintéressemen*t (2009), *Alexis de Tocqueville: The First Social Scientist* (2009), *L'irrationalité* (2010), and *Securities Against Misrule* (2013). They also reflect a deeper immersion in Seneca, Tocqueville, Bentham, and Proust, as well as a belated first reading of *The Theory of Moral Sentiments*, Hume's *History of England*, and Gibbon's *Decline and Fall of the Roman Empire.* A number of books on the American war in Vietnam opened my eyes to the importance of stupidity, however intelligent, in human affairs.

In revising the book, I have given free rein to associations and digressions. My role models in this respect are Montaigne's *Essays*, *The Psychology of Interpersonal Relations* by Fritz Heider, *The Strategy of Conflict* by Thomas Schelling, and *The New Rhetoric* by Chaim Perelman and Lucie-Olbrechts-Tyteca. However different in substance, these books have in common a playful obsession with revealing details, even seemingly trivial ones, superimposed on the analytical structure.

A distinctive feature of this edition will appear when Chapter 9 is read in conjunction with the Conclusion. One might call it the *naturalization of social scientists,* in the sense that I understand many writings by social scientists as instances of the kind of spurious pattern seeking that both natural and social scientists have found to characterize human beings much more generally. I cannot stress enough that this explanation of their explanations is not intended to refute them (this would amount to "the genetic fallacy"). Refutations must follow standard methodological procedures. Yet I believe that the sheer mass of substandard social science – what I call soft and hard obscurantism – calls for an explanation.

My quotations from Proust are taken from the translation by Scott-Moncrieff, occasionally modified either for a more literal rendering or for greater transparency. I thank Herbert Gintis, Aanund Hylland, Yuen Foong Khong, George Loewenstein, Karl Ove Moene, David Stasavage, Adrian Vermeule and Adam Waytz for their comments on an earlier draft.

I dedicate the revised edition to the memory of Aaron Swartz, for his commitment to the public good.

*Part I*

# Explanation and Mechanisms

This book relies on a specific view about explanation in the social sciences. Although not primarily a work of philosophy of social science, it draws upon and advocates certain methodological ideas about how to explain social phenomena. In the first three chapters, these ideas are set out explicitly. In the rest of the book they mostly form part of the implicit background, although from time to time, notably in the Conclusion, they return to the center of the stage.

I argue that all explanation is causal. To explain a phenomenon (an *explanandum*) is to cite an earlier phenomenon (the *explanans*) that caused it. When advocating causal explanation, I do not intend to exclude the possibility of intentional explanation of behavior. Intentions can serve as causes. A particular variety of intentional explanation is *rational-choice explanation*, which will be extensively discussed in later chapters. Many intentional explanations, however, rest on the assumption that agents are, in one way or another, *irrational*.[1] In itself, irrationality is just a negative or residual idea, everything that is not rational. For the idea to have any explanatory purchase, we need to appeal to specific forms of irrationality with specific implications for behavior. In Chapter 14, for instance, I enumerate and illustrate eleven mechanisms that can generate irrational behavior.

Sometimes, scientists explain phenomena by their *consequences* rather than by their causes. They might say, for instance, that blood feuds are explained by the fact that they keep populations down at sustainable levels. This might seem a metaphysical impossibility: how can the existence or occurrence of something at one point in time be explained by something that has not yet come into existence? As we shall see in Chapter 11, the problem can be restated so as to make explanation by consequences a meaningful concept. In the biological sciences, evolutionary explanation offers an example. In the social sciences,

[1] At this first occurrence in the book of the word "agent" it may be worthwhile to note that many scholars prefer "actor." Perhaps economists think in terms of agents, sociologists in terms of actors. Although it does not really matter which term we use, I prefer "agent" because it suggests agency; "actor," by contrast, suggests an audience that may or may not be present.

however, successful instances of such explanations are few and far between. The blood-feud example is definitely not one of them.

The natural sciences, especially physics and chemistry, offer *explanations by law*. Laws are general propositions that allow us to infer the truth of one statement at one time from the truth of another statement at some earlier time. Thus when we know the positions and the velocity of the planets at one time, the laws of planetary motion enable us to deduce and predict their positions at any later (or earlier) time. This kind of explanation is *deterministic*: given the antecedents, only one consequent (or antecedent) is possible. The social sciences offer few if any law-like explanations of this kind. The relation between explanans and explanandum is not one-one or many-one, but one-many or many-many. Many social scientists try to model this relation by using *statistical* methods. Statistical explanations are incomplete by themselves, however, since they ultimately have to rely on intuitions about plausible causal *mechanisms*.

# 1 Explanation

## Explanation: general

The main task of the social sciences is to explain social phenomena. It is not the only task, but it is the most important one, to which others are subordinated or on which they depend. The basic type of explanandum is an *event*. To explain it is to give an account of why it happened, by citing an *earlier event* as its cause. Thus we may explain Ronald Reagan's victory in the 1980 presidential elections by Jimmy Carter's failed attempt to rescue the Americans held hostage in Iran.[1] Or we might explain the outbreak of World War II by citing any number of earlier events, from the Munich agreement to the signing of the Versailles Treaty. Even though in both cases the fine structure of the causal explanation will obviously be more complex, they do embody the basic *event-event* pattern of explanation. In a tradition originating with David Hume, it is often referred to as the "billiard-ball" model of causal explanation. One event, ball A hitting ball B, is the cause of – and thus explains – another event, namely, ball B's beginning to move.

Those who are familiar with the typical kind of explanation in the social sciences may not recognize this pattern, or not see it as privileged. In one way or another, social scientists tend to put more emphasis on *facts*, or states of affairs, than on events. The sentence "At 9 A.M. the road was slippery" states a fact. The sentence "At 9 A.M. the car went off the road" states an event. As this example suggests, one might offer a *fact-event* explanation to account for a car accident.[2] Conversely, one might propose an *event-fact* explanation to account for a given state of affairs, as when asserting that the attack on the World Trade Center in 2001 explains the pervasive state of fear of many Americans. Finally, standard social-science explanations often have a *fact-fact* pattern. To take an

[1] To anticipate a distinction discussed later, note that Carter did not *fail to attempt* but *attempted and failed*. A non-action such as a failure to attempt cannot have causal efficacy, except in the indirect sense that if others perceive or infer that the agent fails to act, they may take actions that they otherwise would not have or decide not to act when they otherwise would have acted.

[2] The voter turnout example discussed later provides another illustration.

example at random, it has been claimed that the level of education of women explains per capita income in the developing world.

Let us consider the explanation of one particular fact, that 65 percent of Americans favor, or say that they favor, the death penalty.[3] In principle, this issue can be restated in terms of events: How did these Americans *come to favor* the death penalty? What were the formative events – interactions with parents, peers, or teachers – that caused this attitude to emerge? In practice, social scientists are usually not interested in this question. Rather than trying to explain a brute statistic of this kind, they want to understand *changes* in attitudes over time or *differences* in attitudes across populations. The reason, perhaps, is that they do not think the brute fact very informative. If one asks whether 65 percent is much or little, the obvious retort is, "Compared to what?" Compared to the attitudes of Americans around 1990, when about 80 percent favored the death penalty, it is a low number. Compared to the attitudes in some European countries, it is a high number.

*Longitudinal* studies consider variations over time in the dependent variable. *Cross-sectional* studies consider variations across populations. In either case, the explanandum is transformed. Rather than trying to explain the phenomenon "in and of itself," we try to explain how it varies in time or space. The success of an explanation is measured, in part, by how much of the variation it can account for.[4] Complete success would explain all observed variation. In a cross-national study we might find, for instance, that the percentage of individuals favoring the death penalty was strictly proportional to the number of homicides per 100,000 inhabitants. Although this finding would provide *no* explanation of the absolute numbers, it would offer a *perfect* explanation of the difference among them.[5] In practice, of course, perfect success is never achieved, but the same point holds. Explanations of variation do not say anything about the explanandum "in and of itself."

An example may be taken from the study of voting behavior. As we shall see later (Chapter 14), it is not clear why voters bother to vote at all in national elections, when it is morally certain that a single vote will make no difference. Yet a substantial fraction of the electorate do turn out on voting day. Why do they bother? Instead of trying to solve this mystery, empirical social scientists usually address a different question: Why does turnout vary across elections? One hypothesis is that voters are less likely to turn out in inclement weather, because rain or cold makes it more attractive to stay home. If the data match

[3] Answers fluctuate. Also, the number of people who favor the death penalty for murder goes down drastically when life imprisonment without parole is stated as the alternative.

[4] Economists sometimes say that they are interested only in what happens "at the margin."

[5] Strictly speaking, the causal chain might go in the other direction, from attitudes to behavior, but in this case that hypothesis is implausible.

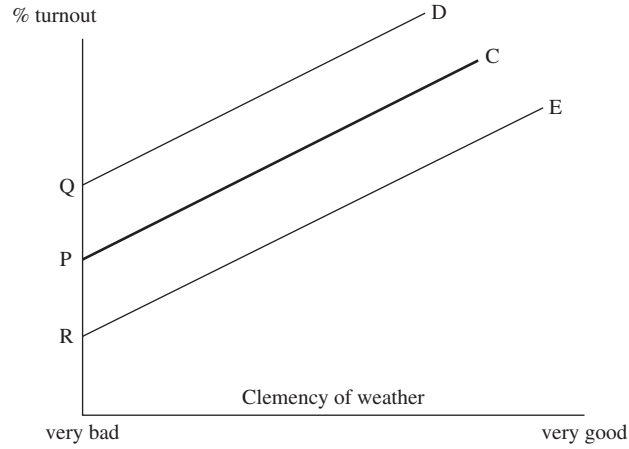



Figure 1.1

this hypothesis, as indicated by line C in Figure 1.1, one might claim to have explained (at least part of) the variation in turnout. Yet one would not have offered *any* explanation of why the line C intersects the vertical axis at P rather than at Q or R. It is as if one took the first decimal as given and focused on explaining the second. For predictive purposes, this might be all one needs. For explanatory purposes, it is unsatisfactory. The "brute event" that 45 percent or more of the electorate usually turn out to vote *is* an interesting one, which cries out for an explanation. I discuss it in several later chapters.

The ideal procedure, in an event-event perspective, would be the following. Consider two elections, A and B. For each of them, identify the events that cause a given percentage of voters to turn out. Once we have thus explained the turnout in election A and the turnout in election B, the explanation of the difference (if any) follows automatically, as a by-product. As a bonus, we might also be able to explain whether identical turnouts in A and B are accidental, that is, due to differences that exactly offset each other, or not. In practice, this procedure might be too demanding. The data or the available theories might not allow us to explain the phenomena "in and of themselves." We should be aware, however, that if we do resort to explanations of variation, we are engaging in a second-best explanatory practice.

Sometimes, social scientists try to explain *non-events*. Why do many people fail to claim social benefits they are entitled to? Why did nobody call the police in the Kitty Genovese case?[6] Considering the first question, the explanation

[6] The version of this episode that has entered the literature is the following. For more than half an hour on March 27, 1964, thirty-eight respectable, law-abiding citizens in Queens, New York,

might be that the individuals in question *decide* not to claim their benefits, because of fear of stigma or concerns with self-image. Since making a decision *is* an event, this would provide a fully satisfactory account. If it fails, social scientists would, once again, look at the *differences* between those who are entitled to benefits and claim them and those who are and do not. Suppose the only difference is that the latter are unaware of their entitlement. As an explanation, this is helpful but insufficient. To go beyond it, we would want to explain *why* some entitled individuals are unaware of their entitlement. To discover that because they are illiterate, they are unable to read the letters informing them about their rights would also be helpful but insufficient. At some point in the explanatory regress, we must either come to a positive event, such as a conscious decision not to become literate or a conscious decision by officials to withhold information, or turn to those who do seek the benefits to which they are entitled. Once we have explained the behavior of the latter, the explanation why others fail to seek their benefit will emerge as a by-product.

Considering the Kitty Genovese case, there is no variation in behavior to explain, since *nobody* called the police. Some accounts of the case indicate that several of the observers *decided* not to call the police. In terms of proximate causes this provides a fully satisfactory account, although we might want to know the reasons for their decision. Was it because they feared "getting involved" or because each observer assumed that someone else would call the police ("Too many shepherds make a poor guard")? Some of the observers, however, apparently did not even think about calling the police. One man and his wife watched the episode for its entertainment value, while another man said he was tired and went to bed. To explain why they did not react more strongly one might cite their shallow emotions, but that, too, would be to account for a negative explanandum by citing a negative explanans. Once again, their behavior can only be explained as a by-product or residual. If we have a satisfactory explanation of why some individuals thought about calling the police, even if in the end they decided not to, we shall have the only explanation we are likely to get of why some did not even think about it.

In the rest of this book I shall often relax this purist or rigorist approach of what counts as a relevant explanandum and an appropriate explanation. The

watched a killer stalk and stab a woman in three separate attacks in Kew Gardens. Twice their chatter and the sudden glow of their bedroom lights interrupted him and frightened him off. Each time he returned, sought her out, and stabbed her again. Not one person telephoned the police during the assault; one witness called after the woman was dead. Although recent research has shown that the version is factually incorrect, the general phenomenon of bystander passivity is well documented (Chapter 12). In references to the case in later chapters I assume the erroneous version, which has become part of the folklore of scholarship. I shall put "Kitty Genovese" in quotation marks, however, to remind the reader that it is a proxy for a more general and better documented class of phenomena.

insistence on event-focused explanations is a bit like the principle of methodological individualism, which is another premise of the book. In principle, explanations in the social sciences should refer only to individuals and their actions. In practice, social scientists nevertheless refer to supraindividual entities such as households, firms, or nations, either as a *harmless shorthand* or as a *second-best approach* forced upon them by lack of data or of fine-grained theories.[7] These two justifications also apply to the use of facts as explananda or as explanantia, to explanations of variation rather than of the phenomena "in and of themselves," and to the analysis of negative explananda (non-events or non-facts). The purpose of the preceding discussion is not to hold social scientists to pointless or impossible standards, but to argue that at the level of first principles the event-based approach is intrinsically superior. If scholars keep that fact in mind they may, at least sometimes, come up with better and more fruitful explanations. When we try to explain the decisions made at the Federal Convention of 1787, the recorded votes of the state *delegations* are useful, but incomplete. Historians have improved our understanding by identifying the votes cast by individual *members* of these delegations. Explanations of why the German National Assembly in 1933 and the French National Assembly in 1940 abdicated their powers gain much in power and focus when we can trace the changing and interacting motivations of individual deputies.

Sometimes, methodological individualism should force us to lower our sights. Social scientists are naturally drawn to big questions, yet some questions may be too big to allow for an answer. We may be able to explain the rise of Calvinism, but not the existence of some form of religion in virtually all societies. We may be able to explain the emergence of capitalist forms of agriculture in eighteenth-century England, but not the "transition from feudalism to capitalism" in Europe as a whole. Discussions of "the Axial age" and "modernity" also flounder, among other reasons, for lack of identifiable agents and their motivations. If social scientists are enjoined to use the microscope rather than the telescope, some questions may of course elude them forever. The loss in breadth is offset, or more than offset, by the gain in depth.

[7] Two economists, correctly observing that "neo-classical utility theory applies to individuals and not to households," set out to explain consumer behavior by appealing only to the preferences of individuals instead of the traditional household-centered approach. Nevertheless, they assume that family decisions are Pareto-efficient, implying that bargaining never breaks down. In real households, however, wives and husbands or parents and children often fail to reach Pareto-efficient decisions, because they do not agree on the division of the jointly created surplus. I mention this not as an objection to their work, which does indeed go beyond the traditional models, but to show that it can be difficult to apply methodological individualism in the absolutely literal sense.

Sometimes, we might want to explain an event (or rather a pattern of events) by its consequences rather than by its causes. I do not have in mind explanation by *intended* consequences, since intentions exist prior to the choices or actions they explain. Rather, the idea is that events may be explained by their *actual* consequences, typically, their *beneficial* consequences for someone or something. As a cause must precede its effect, this idea might seem to be incompatible with causal explanation. Yet causal explanation can also take the form of explanation by consequences, if there is a loop from the consequences back to their causes. A child may initially cry simply because it feels pain, but if the crying also gets it attention from the parents, it may start crying more than it would have done otherwise. I argue in Chapter 11 that this kind of explanation is somewhat marginal in the study of human behavior. In most of the book, I shall be concerned with the simple variety of causal explanation in which the explanans – which might include beliefs and intentions oriented toward the future – precedes the occurrence of the explanandum.[8]

In addition to the fully respectable form of functional explanation that rests on specific feedback mechanisms, there are more disreputable forms that simply point to the production of consequences that are beneficial in some respect and then without further argument assume that these suffice to explain the behavior that causes them. When the explanandum is a *token*, such as a single action or event, this kind of explanation fails for purely metaphysical reasons. To take an example from biology, we cannot explain the occurrence of a neutral or harmful mutation by observing that it was a necessary condition for a further, advantageous one. In a rare moment of methodological sobriety, Marx refers to the speculative distortions by which "later history is made the goal of earlier history, e.g. the goal ascribed to the discovery of America is to further the eruption of the French Revolution." In a less sober moment, he wrote that "The anatomy of man is the key to the anatomy of the ape."

When the explanandum is a *type*, such as a recurrent pattern of behavior, it may or may not be valid. Yet as long as it is not supported by a specific feedback mechanism, we should treat it as if it were invalid. Anthropologists have argued, for instance, that revenge behavior has beneficial consequences of various kinds, ranging from population control to decentralized norm enforcement (Chapter 21 offers many other examples). Assuming that these benefits are in fact produced, they might still obtain by accident. To show that they arise non-accidentally, that is, that they sustain the revenge behavior that causes them, the demonstration of a feedback mechanism is indispensable.

[8] For some purposes, it may be useful to distinguish among causal, intentional, and functional explanation. Physics employs only causal explanation; biology additionally admits functional explanation; and the social sciences further admit intentional explanation. At the most fundamental level, though, all explanation is causal.

And even when one is provided, the initial occurrence of the explanandum must be due to something else.

## The structure of explanations

Let me now turn to a more detailed account of explanation in the social sciences (and, to some extent, more generally). The first step is easily overlooked: before we try to explain a fact or an event we have to establish that the fact *is* a fact or that the event actually did take place. As Montaigne wrote, "I realize that if you ask people to account for 'facts,' they usually spend more time finding reasons for them than finding out whether they are true . . . They skip over the facts but carefully deduce inferences. They normally begin thus: 'How does this come about?' But does it do so? That is what they ought to be asking."

Thus before trying to explain, say, why there are more suicides in one country than in another, we have to make sure that the latter does not tend, perhaps for religious reasons, to underreport suicides. Before we try to explain why Spain has a higher unemployment rate than France, we have to make sure that the reported differences are not due to different definitions of unemployment or to the presence of a large underground economy in Spain. If we want to explain why youth unemployment is higher in France than in the United Kingdom, we need to decide whether the explanandum is the rate of unemployment among young people who are actively searching for jobs or the rate among young people overall, including students. If we compare unemployment in Europe and the United States, we have to decide whether the explanandum is the unemployed in the literal sense, which includes the incarcerated population, or in the technical sense, which only includes those searching for work.[9] Before we try to explain why revenge takes the form of "tit for tat" (I or one of mine kill you or one of yours each time you or yours kill one of mine), we should verify that this is actually what we observe rather than, say, "two tits for a tat" (I kill two of yours each time you or yours kill one of mine). Much of science, including social science, tries to explain things we all know, but science can also make a contribution by establishing that some of the things we think we know simply are not so. In that case, social science may also try to explain *why* we think we know things that are not so, adding as it were a piece of knowledge to replace the one that has been taken away.[10]

[9] In either of the last two cases, some individuals may take up a career as criminals or students because they do not think they would get a job if they tried. For some purposes, one might want to count these among the unemployed; for other purposes, not.

[10] Just as science can help explain popular beliefs in non-facts, it can help explain popular beliefs in false explanations. For instance, most of those who suffer from arthritis believe arthritic pain

Suppose now that we have a well-established explanandum for which there is no well-established explanation – a *puzzle*. The puzzle may be a surprising or counterintuitive fact, or simply an unexplained correlation. One small-scale example is "Why are more theology books stolen from Oxford libraries than books on other subjects?" Another small-scale example, which I shall explore in more detail shortly, is "Why do more Broadway shows receive standing ovations today than twenty years ago?"

Ideally, explanatory puzzles should be addressed in the five-step sequence spelled out in the following. In practice, however, steps (1), (2), and (3) often occur in a different order. We may play around with different hypotheses until one of them emerges as the most promising, and then look around for a theory that would justify it. If steps (4) and (5) are carried out properly, we may still have a high level of confidence in the preferred hypothesis. Yet for reasons I discuss in the next chapter, scholars might want to limit their freedom to pick and choose among hypotheses.

1. Choose the theory – a set of interrelated causal propositions – that holds out the greatest promise of a successful explanation.
2. Specify a hypothesis that applies the theory to the puzzle, in the sense that the explanandum follows logically from the hypothesis.
3. Identify or imagine plausible accounts that might provide alternative explanations, also in the sense that the explanandum follows logically from each of them.
4. For each of these rival accounts, refute it by pointing to additional testable implications that are in fact *not* observed.
5. Strengthen the proposed hypothesis by showing that it has additional testable implications, preferably of "novel facts," that are in fact observed.

These procedures define the *hypothetico-deductive method*. In a given case, they might take the form shown in Figure 1.2. I shall illustrate it by the puzzle of increasing frequency of standing ovations on Broadway. It is not based on systematic observations or controlled experiments, but on my casual impressions confirmed by newspaper reports. For the present purposes, however, the shaky status of the explanandum does not matter. If there are in fact more standing ovations on Broadway than there were twenty years ago, how could we go about explaining it?

is triggered by bad weather. Studies suggest, however, that there is no such connection. Perhaps we should drop the search for the causal link between bad weather and arthritic pain and instead try to explain why arthritics believe there is one. Most likely they were once told there was a connection and subsequently paid more attention to instances that confirmed the belief than to those that did not.

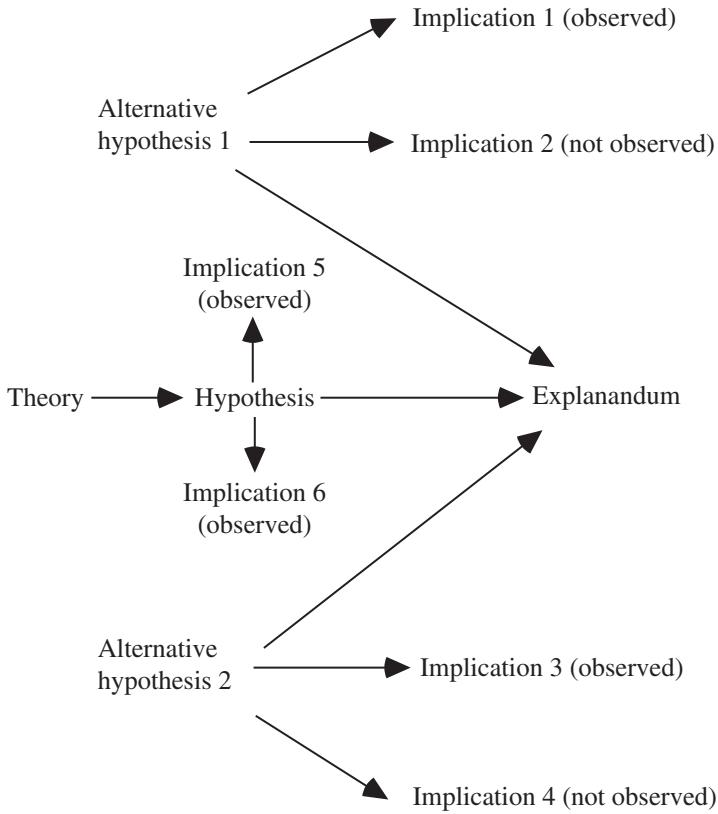



Figure 1.2

I shall consider an explanation in terms of the rising prices of Broadway tickets. One newspaper reports the playwright Arthur Miller as saying, "I guess the audience just feels having paid $75 to sit down, it's their time to stand up. I don't mean to be a cynic but it probably all changed when the price went up." When people have to pay $75 or more for a seat, many cannot admit to themselves that the show was poor or mediocre, and that they have wasted their money. To confirm to themselves that they had a good time, they applaud wildly.

More formally, the explanation is sought in the hypothesis "When people have paid a great deal of money or effort to obtain a good, they tend (other things being equal) to value it more highly than when they paid less for it."[11]

[11] A similar idea is sometimes used to defend the high fees of psychotherapists: patients would not believe in the therapy unless they paid a lot for it. But no therapists to my knowledge state that they donate 50 percent of their fee to Red Cross.

As Montaigne wrote, "where our expenditure is concerned we are good at keeping accounts: our outgoings cost us so much trouble, and we value them precisely because they do so; our opinion will never allow itself to be undervalued. What gives value to a diamond is its cost; to virtue, its difficulty; to penance, its suffering; to medicines, their bitter taste." Given the factual premise of rising prices, this hypothesis passes the minimal test that any proposed explanation must satisfy: If it is true, we can infer the explanandum. But this is a truly minimal test, which many propositions could pass.[12] To strengthen our belief in this particular explanation, we must show that it is supported from below, from above, and laterally.

An explanation is supported *from below* if we can deduce and verify observable facts from the hypothesis over and above the fact that the hypothesis is intended to explain. It must have "excess explanatory power." In the case of the Broadway shows, we would expect fewer standing ovations in shows whose prices for some reason have not gone up.[13] Also, we would expect fewer standing ovations if large numbers of tickets to a show are sold to firms and given by them to their employees. (This would count as a "novel fact.") Even if these tickets are expensive, the spectators have not paid for them out of their own pocket and hence do not need to tell themselves that they are getting their money's worth.

An explanation is supported *from above* if the explanatory hypothesis can be deduced from a more general theory.[14] In the present case, the explanatory proposition is a specification of the theory of cognitive dissonance proposed by Leon Festinger. The theory says that when a person experiences an internal inconsistency or dissonance among her beliefs and values, we can expect some kind of mental readjustment that will eliminate or reduce the dissonance. Typically, the adjustment will choose the path of least resistance. A person who has spent $75 to see a show that turns out to be bad cannot easily make herself believe that she paid less than that amount. It is easier to persuade

[12] The human mind seems to have a tendency to turn this minimal requirement into a sufficient one. Once we have hit upon an account that *may* be true, we often do not pause to test it further or to consider alternative accounts. The choice of an account may be due to the idea of *post hoc ergo propter hoc* (after it, therefore because of it), or to an inference from the fact that a given account is *more plausible than others* to the conclusion that it is *more likely than not* to be correct. Summarizing Jefferson's objection to Voltaire's explanation of the production of seashells, two recent authors write that "science would progress better from honestly recognizing its ignorance . . . than from accepting the most reasonable [among several] far-fetched views."

[13] We would *not* necessarily expect fewer people to rise to their feet in the cheaper sections. They might feel foolish sitting when others are rising; also, they might have to get up to see the actors who would otherwise be blocked from view by those standing in front of them.

[14] More accurately: if it is a *specification* of a more general theory. The relation between a general theory and a specific explanatory hypothesis is rarely a deductive one. For one thing, there may be some slack in the theory itself (see Chapter 2). For another, a given theory can usually be operationalized in many different ways.

herself that the show was in fact quite good. Any show is likely to be good in *some* respect, and by emphasizing that dimension over others spectators can enhance their overall appreciation.

Although not without problems, the theory of cognitive dissonance is pretty well supported. Some of the support is from cases that are very different from the one we are considering here, as when a person who has just bought a car avidly seeks out ads for that very brand of car, to bolster his conviction that he made a good decision. Some of the support arises from quite similar cases, as when the painful and humiliating initiation rituals of college fraternities and sororities induce strong feelings of loyalty. I am not saying that people would consciously tell themselves, "Because I suffered so much to join this group, it must be a good group to belong to." The mechanism by which the suffering induces loyalty must be an unconscious one.

An explanation receives *lateral support* if we can think of and then refute alternative explanations that also pass the minimal test. Perhaps there are more standing ovations because today's audiences, arriving in busloads from New Jersey, are less sophisticated than the traditional audience of blasé New York denizens. Or perhaps it is because shows are better than they used to be. For each of these alternatives, we must think of and then disconfirm additional facts that would obtain if they were correct. If standing ovations are more frequent because audiences are more impressionable, we would expect them also to have been frequent in out-of-town performances twenty years ago. If shows are better than they used to be, we would expect this to be reflected in how well they are reviewed and how long they play before folding.

In this procedure, the advocate for the original hypotheses also has to be the devil's advocate. One has consistently to *think against oneself* – to make matters as difficult for oneself as one can. We should select the strongest and most plausible alternative rival explanations, rather than accounts that can easily be refuted. For similar reasons, when seeking to demonstrate the excess explanatory power of the hypothesis, we should try to deduce and confirm implications that are novel, counterintuitive, and as different from the original explanandum as possible. These two criteria – refuting the most plausible alternatives and generating *novel facts* – are decisive for the credibility of an explanation. Support from above helps but can never be decisive. In the long run it is the theory that is supported by the successful explanations it generates, not the other way around. Emilio Segrè, a Nobel Prize winner in physics, said that some winners confer honor on the Prize whereas others derive honor from it. The latter are, however, parasitic on the former. Similarly, a theory is parasitic on the number of successful explanations it generates. If it is able to confer support on a given explanation, it is only because it has received support from earlier explanations.

### What explanation is not

Statements that purport to explain an event must be distinguished from *seven other types of statement*.

First, causal explanations must be distinguished from *true causal statements*. To cite a cause is not enough: the causal mechanism must also be provided, or at least suggested. In everyday language, in good novels, in good historical writings, and in many social scientific analyses, the mechanism is not explicitly cited. Instead, it is suggested by the way in which the cause is described. Any given event can be described in many ways. In (good) narrative explanations, it is tacitly presupposed that only causally relevant features of the event are used to identify it. If told that a person died as a result of having eaten rotten food, we assume that the mechanism was food poisoning. If told that he died as a result of eating food to which he was allergic, we assume that the mechanism was an allergic reaction. Suppose now that he actually died because of food poisoning, but that he was also allergic to the food in question, lobster. To say that he died because he ate food to which he had an allergy would be true, but misleading. To say that he died because he ate lobster would be true, but uninformative. It would suggest no causal mechanism at all and be consistent with many, such as that he was killed by someone who had taken an oath to kill the next lobster eater he observed.

Second, causal explanations must be distinguished from statements about *correlations*. Sometimes, we are in a position to say that an event of a certain type is invariably or usually followed by an event of another kind. This does not allow us to say that events of the first type cause events of the second, because there is another possibility: the two might be common effects of a third event. In his *Life of Johnson*, Boswell reports that a certain Macaulay, although "with a prejudice against prejudice," affirmed that when a ship arrived at St. Kilda in the Hebrides, "all the inhabitants are seized with a cold." While some offered a causal explanation of this (alleged) fact, a correspondent of Boswell's informed him that "the situation of St. Kilda renders a North-East Wind indispensable before a stranger can land. The wind, not the stranger, occasions an epidemick cold." Or consider the finding that children in contested custody cases are more disturbed than children whose parents have reached a private custody agreement. It could be that the custody dispute itself explains the difference, by causing pain and guilt in the children. It could also be, however, that custody disputes are more likely to occur when the parents are bitterly hostile toward each other and that children of two such parents tend to be disturbed. To distinguish between the two interpretations, we would have to measure suffering before and after the divorce. A third possibility is canvassed later.

Here is a more complex example, my favorite example, in fact, of this kind of ambiguity. In *Democracy in America*, Alexis de Tocqueville discussed the alleged causal connection between marrying for love and having an unhappy marriage. He points out that this connection obtains only in societies in which such marriages are the exception and arranged marriages the rule. Only stubborn people will go against the current, and two stubborn persons are not likely to have a very happy marriage.[15] In addition, people who go against the current are treated badly by their more conformist peers, inducing bitterness and unhappiness. Of these arguments, the first rests on a non-causal correlation, due to a "third factor," between marrying for love and unhappiness. The second points to a true causal connection, but not the one that the critics of love marriages to whom Tocqueville addressed his argument had in mind. Marrying for love causes unhappiness only in a context where this practice is exceptional. Biologists often refer to such effects as "frequency dependent."[16]

In addition to the "third-factor" problem, correlation may leave us uncertain about the *direction* of causality. Consider an old joke:

PSYCHOLOGIST: You should be kind to Johnny. He comes from a broken home.
TEACHER: I'm not surprised. Johnny could break any home.

Or as the comedian Sam Levinson said, "Insanity is hereditary. You can get it from your children." The implication is that a disturbed child may cause the parents to divorce rather than that a divorce causes the disturbance. Similarly, a negative correlation between how much the parents know about what their adolescent children are doing and the children's tendency to get into trouble need not show that parental monitoring works, but only that teenagers intent on getting into trouble are unlikely to keep their parents informed about what they are doing.

Life under Stalin often exhibited such reverse causality. The caption of a cartoon in the satirical *Krokodil* magazine was a brief dialogue: "How come, friend, that you are so often ill?" "I know a doctor" was the answer – not because the doctor made him ill, but because he could issue a much

[15] Here the "third factor" is a character trait, stubbornness, rather than an event.

[16] The first mechanism is a *selection effect*, the second a genuine *aftereffect*. The distinction applies quite widely. If we ask why someone in a certain state (e.g. being in a certain occupation, being unemployed, or being hospitalized for mental illness) is more likely to remain in that state the longer she has already been there, either mechanism (or both) might be at work. The long-term unemployed, for instance, might form a subset of the population with skills for which there is little demand; alternatively, all employed individuals might be equally likely to lose their jobs, but once they lose them, the state of being unemployed changes them (or the perception of them by employers) so that their likelihood of reentering the labor market declines over time. The "labeling theory" of mental illness or crime rests on the (dubious) assumption that aftereffects dominate selection effects.

sought-after certificate of illness. Another cartoon showed a store manager talking politely to a customer, while the check-out clerk and a woman look on. "He's a courteous man, our store manager," says the clerk. "When he sells cloth, he calls all the customers by name and patronymic." "Does he really know all the customers?" "Of course. If he doesn't know someone, he doesn't sell to them."

Third, causal explanations must be distinguished from statements about *necessitation*. To explain an event is to give an account of why it happened *as* it happened. That it might also have happened in some other way, and would have happened in some other way had it not happened the way it did, does not provide an answer to the same question. Consider a person who suffers from cancer of the pancreas, which is certain to kill her within a year. When the pain becomes unendurable, she kills herself. To *explain* why she died within a certain period, it is pointless to say that she *had to* die in that period because she had cancer.[17] If all we know about the case are the onset of cancer, the limited life span of persons with that type of cancer, and the death of the person, it is plausible to infer that she died because of the cancer. We have the earlier event and a causal mechanism sufficient to bring about the later event. But the mechanism is not necessary: it could be preempted by another. (In the example the preempting cause is itself an effect of the preempted cause, but this need not be the case; she might also die in a car accident.) To find out what actually happened, we need more finely grained knowledge. The quest never ends: right up to the last second, some other cause could preempt the cancer.[18]

Statements about necessitation are sometimes called "structural explanations." Tocqueville's analysis of the French Revolution is an example. In his published book on the topic, he cites a number of events and trends from the fifteenth century to the 1780s and asserts that the revolution, against this background, was "inevitable." By this he probably meant (1) that any number of small or medium-sized events would have been sufficient to trigger it and (2) that it was a virtual certainty that *some* triggering events would occur, although not necessarily the ones that actually did happen or when they happened. We may be able to predict the collapse of a house of cards, but not the particular wind gust that will make it fall. Although Tocqueville left notes for a second volume in which he intended to account for the revolution

[17] James Fitzjames Stephen writes that "the law is perfectly clear that, if by reason of [an] assault [a man] died in the spring of a disease which must have killed him, say, in the summer, the assault was the cause of his death."

[18] Causal preemption should be distinguished from causal overdetermination. The latter is illustrated by a person's being hit simultaneously by two bullets, each of which would have been sufficient to kill her. The former is illustrated by a person's being killed by one bullet, preempting the operation of another fired a few seconds later.

as it *did* happen, one might argue that if he successfully established (1) and (2), there was no need to take this further step. The problem with this line of reasoning is that in many interesting social-science questions (and in contrast to the cancer example), claims such as (1) and (2) are very hard to establish by methods untainted by hindsight.[19] A stronger argument can be made when similar events happen independently of each other at the same time, suggesting that they were "in the air." The study of simultaneous and independent rumors provides an example.

Fourth, causal explanation must be distinguished from *storytelling*. A genuine explanation accounts for what happened, as it happened. To tell a story is to account for what happened as it *might* have happened (and perhaps did happen). I have just argued that scientific explanations differ from accounts of what *had to* happen. I am now saying that they also differ from accounts of what *may* have happened. The point may seem trivial, or strange. Why would anyone want to come up with a purely conjectural account of an event? Is there any place in science for speculations of this sort? The answer is yes – but their place must not be confused with that of explanation.

Storytelling can suggest new, parsimonious explanations. Suppose that someone asserts that self-sacrificing or helping behavior is conclusive proof that not all action is self-interested, and that emotional behavior is conclusive proof that not all action is rational. One might conclude that there are three irreducibly different forms of behavior: rational and selfish, rational and non-selfish, and irrational. The drive for parsimony that characterizes good science should lead us to question this view. Might it not be the case that when people help others it is because they expect reciprocation, and that when they become angry it is because that helps them to get their way? By telling a story about how rational self-interest *might* generate altruistic and emotional behavior, one can transform an issue from a philosophical one into one that is amenable to empirical research.[20] A just-so story can be the first step in the construction of a successful explanation.

At the same time, storytelling can be misleading and harmful if it is mistaken for genuine explanation. With two exceptions stated in the next paragraph, "as-if" explanations do not actually explain anything. Consider for instance the common claim that we can use the rational-choice model to

[19] The American Revolution is perhaps a more plausible candidate for a structural explanation. An acute neutral observer such as the French minister Choiseul observed as early as 1765 that the independence of the American colonies was inevitable. For a detached French commentator such as Raymond Aron, the independence of Algeria was also a foregone conclusion well before it came about. The French Revolution is more akin to the collapse of Communism – inevitable mainly in hindsight.

[20] In this particular case, the just-so stories happen to be false, since people also help others in one-shot interactions and getting angry may cause others to refrain from interacting with them.

explain behavior, even though we *know* that people cannot perform the complex mental calculations embodied in the model (or in the mathematical appendixes of the articles in which the model is set out). As long as the model provides predictions with a good fit with the observed behavior, we are entitled (it is claimed) to assume that agents act "as if" they are rational. This is the operationalist or instrumentalist view of explanation, which originated in physics and was later adopted by Milton Friedman for the social sciences (see Chapter 11). The reason, it is claimed, we can assume that a good billiards player knows the law of physics and can carry out complex calculations in his head is that this assumption enables us to predict and explain his behavior with great accuracy. To ask whether the assumption is *true* is to miss the point.

This argument may be valid in some situations, in which the agents can learn by trial and error over time. It is valid, however, precisely because we can point to a *mechanism* that brings about non-intentionally the same outcome that a superrational agent could have calculated intentionally. In the absence of such a mechanism, we might still accept the instrumentalist view if the assumption enabled us to predict behavior with very great accuracy. The law of gravitation seemed mysterious for a long time, as it seemed to be based on the unintelligible idea of action at a distance. Yet because it made possible predictions that were accurate to many decimal points, Newton's theory was uncontroversially accepted until the advent of the theory of general relativity. The mysterious workings of quantum mechanics are also accepted, albeit not always without qualms, because they allow for predictions with even more incredible accuracy.

Rational-choice social science can rely on neither of these two supports. *There is no general non-intentional mechanism that can simulate or mimic rationality*. Reinforcement learning (Chapter 11) may do it in some cases, although in others it produces systematic deviations from rationality. Some kind of social analog to natural selection might do it in other cases, at least roughly, if the rate of change of the environment is less than the speed of adjustment (Chapter 11). In one-shot situations or in rapidly changing environments, I do not know of any mechanism that would simulate rationality. At the same time, the empirical support for rational-choice explanations of complex phenomena tends to be quite weak. This is of course a sweeping statement. Rather than having to explain what I mean by "weak," let me simply point to the high level of disagreement among competent scholars about the explanatory force of competing hypotheses. Even in economics, in some ways the most developed among the social sciences, there are fundamental, persistent disagreements among "schools." We *never* observe the kind of many-decimal-points precision that would put controversy to rest.

Fifth, causal explanations must be distinguished from *statistical explanations*. Although many explanations in the social sciences have the latter form,

they are unsatisfactory in the sense that they cannot account for individual events. To apply statistical generalizations to individual cases is a grave error, not only in science but also in everyday life.[21] Suppose it is true that men tend to be more aggressive than women. To tell an angry man that his anger is caused by his male hormones rather than argue that it is unjustified by the occasion is to commit both an intellectual and a moral fallacy. The intellectual fallacy is to assume that a generalization valid for most cases is valid in each case.[22] The moral fallacy is to treat an interlocutor as governed by biological mechanisms rather than as open to reason and argument.

Although statistical explanations are always second best, in practice we may not be able to do any better. It is important to note, however, that they are inevitably guided by the first-best ideal of causal explanation. It appears to be a statistical fact that citizens in democracies live longer than citizens in non-democratic regimes. Before we conclude that the political regime explains longevity, we might want to *control for* other variables that might be responsible for the outcome. It might be that more democracies than non-democracies have property X, and that it is really X that is responsible for life expectancy. But as there are indefinitely many such properties, how do we know which to control for? The obvious answer is that we need to be guided by a causal hypothesis. It seems plausible, for instance, that citizens in industrialized societies might live longer than citizens of less developed societies. If industrial societies also tend to be more democratic than non-industrial regimes, that could account for the observed facts. To make sure that it is democracy rather than industrialization that is the causal factor, we have to compare democratic and non-democratic regimes at the same level of industrialization, and see whether a difference persists. Once we feel reasonably confident that we have controlled for other plausible causes, we may also try to find out *how* – by which causal chain or mechanism – the regime type affects life span. I discuss this second step in the next chapter. Here, I only want to note that our confidence is inevitably based on *causal intuitions* about what are (and what are not) plausible "third factors" for which we need to control.[23]

[21] The converse fallacy – using an individual case to generate or support a generalization – is equally to be avoided. Proust wrote that the housekeeper Françoise in the Narrator's family "was as likely to take the particular for the general as the general for the particular." This combination can be pernicious. Suppose you observe a member of group X telling a lie. Generalizing, you form the belief that members of group X tend to lie. Next, observing another member of the group, you assume he is lying. Finally, the (unverified) assumption is used as further evidence for the generalization.

[22] An example is the recent American practice of "evidence-based sentencing," where the evidence refers not to the particular case but to statistics about the risk of recidivism for members of the *group* to which the defendant belongs.

[23] For instance, there is no plausible causal mechanism that should make us control for the population size of democratic and non-democratic regimes. Although one cannot exclude a

Sixth, causal explanations must be distinguished from why-explanations, that is, answers to *"why-questions."* Suppose we read a scholarly article and see to our surprise that the author does not refer to an important and relevant article, causing us to ask ourselves, "Why does he not cite it?" Our curiosity may be perfectly satisfied if we learn that he was in fact unaware of that earlier work (although we might also want to know why he had not explored the literature more thoroughly). But "He did not cite it because he was not aware of it" is not a causal explanation. If read as a causal explanation it would imply, absurdly, citing a non-event to explain another non-event. ("The reason they never married is that they never met.") Suppose, however, that we discover that the author was aware of the article but *decided* not to cite it because he himself had not been cited in it. In that case the answer to the why-question also provides a causal explanation. There is an event, the decision to not cite the article, caused by an earlier event, the anger triggered by not being cited.

Although why-explanations of non-events do not provide a causal account, they are perfectly respectable. They satisfy our curiosity, and substitute understanding for puzzlement. I pursue the question in Chapter 10.

Finally, causal explanations must be distinguished from *predictions*. Sometimes we can explain without being able to predict, and sometimes predict without being able to explain. True, in many cases one and the same theory will enable us to do both, but I believe that in the social sciences this is the exception rather than the rule.

I postpone the main discussion of why we can have explanatory power without strong predictive power to the next chapter. In brief preview, the reason is that in many cases we can identify a causal mechanism after the fact, but not predict before the fact which of several possible mechanisms will be triggered. The special case of biological explanation is somewhat different. As further discussed in Chapter 11, evolution is fueled by the twin mechanisms of random mutations and (more or less) deterministic selection. Given some feature or behavioral pattern of an organism, we can explain its *origin* by appealing to a random change in the genetic material and its *persistence* by its favorable impact on reproductive fitness. Yet prior to the occurrence of the mutation, no one could have predicted it. Moreover, as the occurrence of one mutation constrains the subsequent mutations that can occur, we may not even be able to predict that a given mutation will occur sooner or later. Hence structural explanations are unlikely to be successful in biology. The phenomenon of *convergence* – different species' developing similar adaptations because they are under similar environmental pressures – has a structural flavor but does not allow us to say that the adaptations were inevitable.

causal link between population size and average life span, social science has not established any such connection; nor can I imagine a non-contrived one.

Conversely, we may have predictive power without explanatory power. To predict that consumers will buy less of a good when its price goes up, there is no need to form a hypothesis to explain their behavior. Whatever the springs of individual action – rational, traditional, or simply random – we can predict that overall people will buy less of the good simply because they can afford less of it (Chapter 10). Here there are several mechanisms that are constrained to lead to the same outcome, so that for predictive purposes there is no need to choose among them. Yet for explanatory purposes, the mechanism is what matters. It provides understanding, whereas prediction offers at most control.

Also, for predictive purposes the distinction among correlation, necessitation, and explanation becomes pointless. If there is a law-like regularity between one type of event and another, it does not matter – for predictive purposes – whether it is due to a causal relation between them or to their being common effects of a third cause. In either case we can use the occurrence of the first event to predict the occurrence of the second. Nobody believes that the first symptoms of a deadly disease cause the later death, yet they are regularly used to predict that event. Similarly, if knowing a person's medical condition allows us to predict that he will not be alive one year from now, the prediction is not falsified if he dies of a car accident or if he takes his life because the illness is too painful.

### Bibliographical note

The general view of explanation and causation on which I rely is set out in more detail in J. Elster, D. Føllesdal, and L. Walløe, *Rationale Argumentation* (Berlin: Gruyter, 1988). For applications to human action I refer the reader to D. Davidson, *Essays on Actions and Events* (Oxford University Press, 1980). My criticism of functional explanation is set out in various places, notably in *Explaining Technical Change* (Cambridge University Press, 1983). The classical version of the Kitty Genovese case is A. M. Rosenthal, *Thirty-Eight Witnesses* (Berkeley: University of California Press, 1999), corrected by R. Manning, M. Levine, and A. Collins, "The Kitty Genovese murder and the social psychology of helping," *American Psychologist* 62 (2007), 555–62. An outstanding "micro-political" account of the abdication from power by the German and French assemblies is I. Ermakoff, *Ruling Oneself Out* (Duke University Press, 2008). The attempt to provide micro-foundations for consumer behavior is M. Browning and P. A. Chiappori, "Efficient intra-household allocations," *Econometrica* 66 (1998), 1241–78. A convenient access to Festinger's views is in L. Festinger, S. Schachter, and M. Gazzaniga (eds.), *Extending Psychological Frontiers: Selected Works of Leon Festinger* (New York: Russell Sage, 1989). The observation on Jefferson and sea-shells is taken from C. Calomiris and S. Haber, *Fragile by Design*

(Princeton University Press, 2014), p. 480. The examples of "child-to-parent" effects are from two stimulating books by J. R. Harris, *The Nurture Assumption: Why Children Turn Out the Way They Do* (New York: Free Press, 1998) and *No Two Alike* (New York: Norton, 2006). The captions to the *Krokodil* cartoons are cited from S. Fitzpatrick, *Everyday Stalinism* (University of Chicago Press, 1999), p. 65. I discuss Tocqueville's views on causality in "Patterns of causal analysis in Tocqueville's *Democracy in America*," *Rationality and Society* 3 (1991), 277–97, and his views on the French Revolution in "Tocqueville on 1789," in C. Welch (ed.), *The Cambridge Companion to Tocqueville* (Cambridge University Press, 2006). Milton Friedman's defense of "as-if" rationality in "The methodology of positive economics" (1953) is reprinted in M. Brodbeck (ed.), *Readings in the Philosophy of the Social Sciences* (London: Macmillan, 1969). For an empirical criticism of his argument, see T. Allen and C. Carroll, "Individual learning about consumption," *Macroeconomic Dynamics* 5 (2001), 255–71. A defense of the "as-if" approach in political science is R. Morton, *Methods and Models: A Guide to the Empirical Analysis of Formal Models in Political Science* (Cambridge University Press, 1999). Like most other defenders of the approach, she does not offer a reason why we should *believe* in the "as-if" fiction. A partial exception is D. Satz and J. Ferejohn, "Rational choice and social theory," *Journal of Philosophy* 91 (1994), 71–87. The discussion of why-questions draws on B. Hansson, "Why explanations," *Theoria* 72 (2006), 23–59. The independence of the law of demand from motivational assumptions was noted in G. Becker, "Irrational behavior in economic theory," *Journal of Political Economy* 70 (1962), 1–13.

# 2 Mechanisms

### Opening the black box

Philosophers of science often argue that an explanation must rest on a *general law*. To explain an event is to cite a set of initial conditions together with a statement to the effect that whenever those conditions obtain an event of that type follows. In this chapter I offer two objections to this idea, one moderate and relatively uncontroversial, the other more radical and open to dispute.

The first objection is that even if we can establish a general law from which we can deduce the explanandum (the second objection denies that we can always do this), this does not always amount to an explanation. Once again, we may refer to the distinction between explanation on the one hand and correlation and necessitation on the other. A general law to the effect that certain symptoms of a disease are always followed by death may not explain why the person died. A general law based on the fundamental nature of the disease does not explain the death if the disease was preempted by a suicide or a car accident.

To get around these problems, it is often argued that we should replace the idea of a general law with that of a *mechanism*. Actually, as I use the term "mechanism" in a special sense later, I shall use the phrase "causal chain" to denote what I have in mind here.[1] Rather than trying to explain an event E by the statement "Whenever events C1, C2, . . ., Cn occur, an event of type E follows," one may try to establish the causal chain that leads from the causes C1, C2, . . ., Cn, up to E. This step is often referred to as "opening the black box." Suppose we know that heavy smokers are much more likely than others to get lung cancer. This fact might be due to the fact either that smoking is a cause of lung cancer or that people disposed to smoking also are disposed to the cancer (perhaps genes predisposing for lung cancer are linked to genes that make some people more readily addicted to nicotine).[2] To establish the former

[1] In some of my earlier writings I used "mechanism" to denote what I now call "causal chains." In more recent work I began to use "mechanism" in the sense defined later in this chapter. I should probably have chosen a different terminology, but it is too late now.

[2] As noted later, the second explanation was at one point seriously proposed.

explanation, we will have to exhibit a chain of physiological cause-effect relations that begins with heavy smoking and ends with lung cancer. The final explanation will be more fine-grained, have more causal links, and be more convincing than the black-box statement "Smoking causes cancer."

Or suppose that somebody asserted that high unemployment causes wars of aggression and adduced evidence for a law-like connection between the two phenomena. Once again, how can we know that this is a causal effect and not a mere correlation? Perhaps high fertility rates, which cause unemployment, also motivate political leaders to initiate aggressive wars? Unsuccessful wars would at least cut down the population size, and successful ones would provide new territories for expansion and migration. To eliminate this possibility, we would first control for fertility rates (and other plausible "third factors") and see whether the connection remains. If it does, we would still not be satisfied until we are provided with a glimpse inside the black box and told *how* high unemployment causes wars. Is it because unemployment induces political leaders to seek new markets through wars? Or because they believe that unemployment creates social unrest that must be channeled toward an external enemy, to prevent revolutionary movements at home? Or because they believe that the armaments industry can absorb unemployment? Or could it be that the unemployed tend to vote for populist leaders who are likely to eschew diplomacy and instead use wars to resolve conflicts?

Consider the last proposal in more detail. *Why* would the unemployed vote for irresponsible populist leaders rather than for politicians from one of the established parties? Once again, one can imagine a number of ways of opening this particular black box. Perhaps the natural clientele of populist politicians are more likely to vote when they are unemployed, because their opportunity cost of voting (that is, the value of their time) is less than it is when they have a job. Or perhaps populist leaders are more likely to propose instant solutions to the unemployment problem. Or perhaps they offer policies that would punish those whom the unemployed believe to be responsible for their plight or to benefit from it, be they capitalists or an economically successful ethnic minority.

Consider the last proposal in more detail. *Why* would the unemployed want to punish capitalists or affluent minorities? Is that not just another black-box statement? One way of spelling it out would be by asserting that the unemployed are motivated by material self-interest. If the state could confiscate the wealth of these elites, the funds could be used for redistribution to benefit the unemployed. Or perhaps they are motivated by a desire for revenge, which would incite them to punish the elite even if they would not benefit in material terms. If the rich are seen as engaging in ruthless downsizing to increase their profits, those who lose their jobs can use the ballot box to get

even. Or the unemployed might simply be envious of the clever minority members who succeed where they failed and use the ballot box to cut them down to size.

As far as I know, high unemployment does not cause wars of aggression. The whole exercise is hypothetical. Yet I believe it supports the idea that the credibility of an explanation increases with the extent to which general laws are spelled out in terms of a causal chain. At the level of general laws we can never be sure that we have controlled for all relevant "third factors." There may always be some cause lurking in the wings that would account for both the explanandum and its alleged cause. If we increase the number of links in the causal chain, we reduce this danger.

The danger cannot, however, be eliminated. Specifying a causal chain does not mean giving up on general laws altogether, only going from general laws at a high level of abstraction to laws at a lower level of abstraction. We might, for instance, replace the universal law "High unemployment causes wars" by the less abstract laws "Populist leaders are war prone" and "The unemployed vote for populist leaders." The latter law, in turn, might be replaced by the conjunction of "The unemployed are envious of rich minorities" and "Those who envy rich minorities vote for populist leaders." As with any other law, these might turn out to be mere correlations. If being envious of rich minorities and being unemployed are common effects of a joint cause, the electoral success of war-prone leaders would be due not to unemployment but to a factor causally correlated with it. Yet at this more fine-grained level, there are fewer factors to control for. The better we focus the causal story, the easier it is to make sure that we are not dealing with mere correlation.

Explanations in terms of (very) general laws are also unsatisfactory because they are too opaque.[3] Even if presented with an ironclad case for a universal link between unemployment and wars of aggression and a persuasive argument that all remotely plausible "third factors" have been controlled for, we would still want to know *how* unemployment causes wars. We might believe that the explanation is correct, and yet not be satisfied with it. As I noted in the previous chapter, this was the status of explanations relying on the law of gravitation before general relativity. Action at a distance was so mysterious that many refused to believe it could be the last word. As the law allowed for correct predictions to many decimal points, skeptics had to accept that things happened "as if" it were true, although they would not accept the existence of a force that could "act where it was not."

[3] Some mathematicians are unhappy with the computer-generated proof of the four-color theorem, because they do not provide an intuitive understanding of *why* it is true.

## Mechanisms

Readers may well have said to themselves that the instances of alleged universal laws in this exercise are pretty implausible. I agree. In part, their lack of plausibility may be due to the limits of my imagination in concocting the examples, but I believe there are deeper reasons too. There are simply very few well-established general laws in the social sciences. The "law of demand" – when prices go up, consumers buy less – is well supported, but as laws go it is pretty weak.[4] The law of gravitation, for instance, says not only that as the distance between two objects increases the attractive force between them decreases: it tells us by *how much* it decreases (inversely with the square of the distance). There is nothing like the law of gravitation in the social sciences.[5]

The law of demand and Engel's law, according to which the fraction of income used on food declines as income increases, are what we might call *weak laws*. For any change (up or down) in the independent variable they allow us to predict the *direction* or the sign of a change (up or down) in the dependent variable. They do not allow us, however, to predict the *magnitude* of the change. Although weak, such laws have some content, since they allow us to rule out a whole range of possible values of the dependent variable. They do not help us, however, to single out the value that will be realized within the non-excluded range.

The law of demand is not only weak, but also badly suited for explanatory purposes. As we saw in Chapter 1, it is compatible with several assumptions about how consumers behave. To *explain* why consumers buy less of a good when it becomes more expensive we would have to adopt and test a specific assumption about individual consumer reactions to price changes. Specifically, we have to rely on what I call *mechanisms*. Roughly speaking, mechanisms are *frequently occurring and easily recognizable causal patterns that are triggered under generally unknown conditions or with indeterminate consequences*. They allow us to explain, but not to predict. It has been argued, for instance, that for every child who becomes alcoholic in response to an alcoholic environment, another eschews alcohol in response to the same environment. Both reactions embody mechanisms: doing what your parents do and

[4] Moreover, for some goods demand goes *up* when prices go up. Consumers may be attracted to a good because it is expensive (the "Veblen effect") or buy less of a good such as bread when its price falls because they can afford to replace some of it by higher-quality goods such as meat (the "Giffen effect").

[5] To be sure, it is often said that the strength of altruistic feelings toward others varies inversely with their social distance from the agent. Yet the idea of "social distance" is more like a metaphor than like a concept, and in any case "varies inversely" is much less precise than "varies inversely with the square of the distance."

doing the opposite of what they do. We cannot tell ahead of time what will become of the child of an alcoholic, but if he or she turns out either a teetotaler or an alcoholic we may suspect we know why.

I do not claim that there is any kind of objective indeterminacy at work here; indeed that concept has little meaning outside quantum mechanics. I am claiming only that we can often explain behavior by showing it to be an instance of a general causal pattern, even if we cannot explain why that pattern occurred. The mechanisms of conformism (for instance, doing what your parents do) and of anti-conformism (doing the opposite of what they do) are both very general. If we can show the behavior of a child with an alcoholic parent to be an instance of one or the other mechanism, we have provided an explanation of the behavior. One might object that as long as we have not shown why the child became (say) an alcoholic rather than a teetotaler we have not explained anything. I would certainly agree that an account showing why one rather than the other outcome occurred would be a better one, and I do not deny that we might sometimes be able to provide one. But to subsume an individual instance under a more general causal pattern is also to provide an explanation. To know that the child became an alcoholic as a result of conformism is to remove some of the opaqueness of the outcome, although some will remain as long as we do not also explain why the child was subject to conformism.

I said that a mechanism is "a frequently occurring and easily recognizable causal pattern." Proverbial folk wisdom has identified many such patterns.[6] In my preferred definition, "A proverb has been passed down through many generations. It sums up, in one short phrase, a general principle, or common situation, and when you say it, everyone knows exactly what you mean." Moreover, proverbs often state mechanisms (in the sense used here) rather than general laws. Consider, in particular, the striking tendency for proverbs to occur in mutually exclusive pairs. On the one hand, we have "Absence makes the heart grow fonder," but on the other "Out of sight, out of mind." On the one hand we may think that forbidden fruit tastes best, but on the other that the grapes beyond our reach are sour. On the one hand, "Like attracts like," but on the other "Opposites attract each other." On the one hand, "Like father, like son," but on the other "Mean father, prodigal son." On the one hand, "Haste makes waste," but on the other "He who hesitates is lost." On the one hand, "To remember a misfortune is to renew it," but on the other "The remembrance of past perils is pleasant." (As noted later, the last two are in fact not mutually exclusive.) Many other examples could be cited.

[6] As we shall see in Chapter 12, however, proverbs are not always wise.

Many pairs of opposite mechanisms do not appear to be captured by proverbs. Consider for instance what we may call the spillover–compensation pair. If a person who works very hard at the job goes on vacation, would we expect her to carry over the same frenetic pace to her leisure activities (spillover effect) or on the contrary to relax utterly (compensation effect)? Or would we expect citizens in democracies to be prone or averse to religion? If they carry over the habit of deciding for themselves from the political to the religious sphere (spillover), we would expect weak religious beliefs. If the lack of a superior authority in politics leads them to seek authority elsewhere (compensation), a democratic political regime would rather tend to favor religion. A contemporary question, which still seems to be undecided, is whether violence on television stimulates real-life violence (spillover) or attenuates it (compensation).

Similar mechanisms can apply to relations among individuals. Consider the question of explaining donations to charity. One individual may be mainly concerned with the efficiency of giving. If others give little, his donation will make more of an impact and hence he is more likely to give; if others give much his donation matters less and he may not make any. Another donor may be more concerned with fairness (among donors). If others give little, she cannot see why she should give more; conversely if others give much she may feel compelled to follow suit. The same pair of mechanisms may apply in collective-action situations. As a popular movement grows, some individuals may drop out because they do not believe they make much of a difference any more, whereas others may join because they do not feel they should stay on the sidelines while others are paying the cost (Chapter 23).

Or consider a saying by La Fontaine to which I shall return several times in this book, "Everyone believes very easily what they fear and what they desire."[7] Although the statement is implausible if read literally, as a universal law it is a useful reminder that in addition to the well-known phenomenon of wishful thinking, there is a less-well-understood propensity to what we might call *countermotivated thinking*. In *À la recherche du temps perdu*, the Narrator reflects on the similarity between his jealousy toward Albertine and Swann's jealousy toward Odette (for the latter, see Chapter 7):

> I had long since been prepared, by the strong impression made on my imagination and my faculty for emotion by the example of Swann, *to believe in the truth of what I feared rather than of what I should have wished.* And so the comfort brought me by Albertine's affirmations [of being faithful to him] came near to being jeopardized for a moment, because I was reminded of the story of Odette. But I told myself that, if it

[7] If one person fears and another person hopes for the same event, they may converge on the unfounded belief that it will occur, while more sober persons discard it. I give several examples in later chapters of such cognitive alliances of motivational opposites.

was only right to allow for the worst, not only when, in order to understand Swann's sufferings, I had tried to put myself in his place, but now, when I myself was concerned, in seeking the truth as though it referred to some one else, still I must not, out of cruelty to myself, a soldier who chooses the post not where he can be of most use but where he is most exposed, end in *the mistake of regarding one supposition as more true than the rest, simply because it was more painful.*

The first sentence I have italicized is an echo of La Fontaine. The second italicized sentence describes the phenomenon of second-order wishful thinking, the belief that one's counterwishful thinking may be due to *overcorrection of wishful thinking* because of "cruelty to oneself."

Consider, finally, the proverbial claims "Too many shepherds make a poor guard" and "Too many cooks make the soup too salty." Again, the value of the proverbs is not to state a universal law, but to suggest mechanisms. The first proverb might be true if each shepherd believes that everybody else is keeping watch (remember the "Kitty Genovese" case), and the second if each cook believes that nobody else is adding salt to the soup.

Even proverbs that are not matched with an opposite proverb often express mechanisms rather than laws. The proverb "The best swimmers drown" would be absurd if taken to mean that the propensity to drown invariably increases with swimming skill. Yet for some swimmers it may indeed be the case that their confidence in their swimming skill increases more rapidly than their skill, causing them to take unwarranted risks ("Pride goes before a fall"). The proverb "People who listen at doors rarely hear anything favorable about themselves" alerts us to the possibility that people who listen at doors may have other disagreeable (and more observable) behavioral traits, but, as we shall see in Chapter 12, such correlations are far from perfect.

When defining mechanisms, I also said that they "are triggered under generally unknown conditions or with indeterminate consequences." Most of the proverbial mechanisms that I have cited so far fall into the first category. We do not know which conditions will trigger conformism or anti-conformism, wishful thinking or counterwishful (countermotivated) thinking, adaptive preferences (sour grapes) or counteradaptive preferences (the grass is greener). We know that at most one member of each pair will be realized, but we cannot tell which. The qualification "at most" is important, because some people may not be subject to either member of these mechanism pairs. Genuine autonomy means being neither conformist nor anti-conformist. (In fact, many anti-conformists conform to one another.) Some people accept that they may not be able to achieve their highest aims, without seeking peace of mind by denigrating these aims.

In other cases, proverbs suggest the simultaneous triggering of two mechanisms with oppositely directed effects on the outcome. In that case, the indeterminacy lies in determining the *net effect* of the mechanisms rather than

in determining which of them (if any) will be triggered. Consider for instance "Necessity is the mother of invention" and "It is expensive to be poor." The first proverb asserts a causal link between poverty and a strong *desire* for innovation, the second a link between poverty and few *opportunities* for innovation. Because behavior is shaped by desires as well as by opportunities (Chapter 10), we cannot in general tell whether the net impact of poverty on innovation is positive or negative. Or consider the pair of proverbs mentioned earlier, "To remember a misfortune is to renew it" versus "The remembrance of past perils is pleasant." The first proverb relies on what has been called an "endowment effect": the memory of a bad experience is a bad experience.[8] The second relies on a "contrast effect": the memory of a bad experience enhances the value of the present.[9] In general we cannot tell whether the net effect of an early bad experience on later welfare will be positive or negative.

Once again, we are not restricted to proverbs. Consider for instance two non-proverbial mechanisms involved in what has been called "the psychology of tyranny." If the tyrant steps up the oppression of the subjects, two effects are likely to occur. On the one hand, harsher punishments will deter them from resistance or rebellion. On the other hand, the more he behaves as a tyrant the more they will hate him, increasing the likelihood of resistance.

If hatred dominates fear, oppression will backfire. Gibbon wrote that the "wanton and ill-timed cruelty" of the Emperor Maximin "instead of striking terror, inspired hatred." In countries occupied by Germans during World War II, members of the resistance sometimes exploited this mechanism when they killed German soldiers to provoke a reprisal, on the assumption that the deterrence effect would be dominated by the "tyranny effect."[10] After September 11, 2001, the United States learned the truth of Seneca's dictum: "a cruel king increases the number of his enemies by destroying them; for the parents and children of those who are put to death, and their relatives and friends, step into the place of each victim." Echoing this statement, John Paul Vann objected to the American strategy in Vietnam that the bombing and shelling "kills many, many more civilians than it ever does VC [Viet Cong] and as a result makes more VC." In the issue of *Le Monde* dated January 21, 2014 the headline of an article on the uprising in Kiev reads, "The adoption of

[8] Conversely, the memory of a good experience is a good experience. Thus Tennyson: "'Tis better to have loved and lost than never to have loved at all."

[9] Conversely, the memory of a good experience devalues the present. Thus Donne: "'Tis better to be foul than to have been fair."

[10] Sometimes the hatred found a different target. In three villages in Central and Northern Italy where the Germans undertook savage reprisals in 1944, some villagers were still hostile to the partisans fifty years later because they were seen as indirectly or even "truly" responsible for the massacre. When A causes B to kill C, relatives and friends of C may direct their anger at A rather than B. In the resistance movements that were fighting German occupational troops, both mechanisms were observed.

repressive laws that were intended to put an end to the contestation, provokes an escalation."

In other cases the net effect is indeterminate. Commenting on the persecution of heretics under Henry VIII, Hume writes that "those severe executions, which in another disposition of men's minds, would have sufficed to suppress [the new doctrine], now served only to diffuse it the more among the people, and to inspire them with horror against the unrelenting persecutors." The indeterminacy is illustrated in a cartoon from the London *Observer* on January 4, 2009. It shows a young boy on a heap of rubble in Gaza, watching Israeli bombers and asking himself "Is this going to make me more or less likely to fire rockets at Israel when I grow up?"[11]

For another instance of indeterminacy, consider the case of a person who faces a barrier or impediment to her goal. This threat to her freedom of action may induce what psychologists call "reactance" – a motivation to recover or reestablish the freedom. The effects of the barrier and the consequent reactance oppose each other, and in general we cannot tell which will be the stronger.[12] As an illustration, think of the effect of hiding from a small boy a drum his parents do not want him to play with. I return to the mechanism of reactance in Chapter 9.

Even when we know the net effect, we may not be able to explain it. Suppose we were somehow able to observe and measure a zero net effect of the endowment and contrast effects with regard to a good experience in the past. This outcome might come about in two ways. Although the three-star French meal I had last year reduced my pleasure from later meals in more ordinary French restaurants, this negative effect on my welfare is exactly offset by the memory of what a great meal it was. Yet the observation of a zero net effect is also perfectly consistent with both endowment and contrast effects of zero – as well as with both effects being very and equally strong. As long as we do not know which is the case, we cannot claim to have explained the outcome. To assess the strength of each effect, we might look at the outcome in a situation in which the other is not expected to occur. If, as seems plausible, my pleasure from Greek cooking is unaffected by the three-star French meal, we can identify the strength of the pure endowment effect.

[11] In 2014, the Shejaiya Brigade of Hamas allegedly issued a combat manual encouraging its fighters to deploy in densely populated areas and stating that "the destruction of civilian homes" in Gaza by Israeli bombing was welcome because it "increases the hatred of the citizens towards the attackers." Whether or not the allegation is correct, the strategy is not without precedent. It can backfire in two ways, either if the civilians' fear dominates their hatred or if their hatred is directed at those who cause civilians to be killed by the enemy rather than at the enemy (see previous footnote).

[12] A special feature of this example is that one of the two competing effects (the reactance) is induced by the other (the barrier). In the other examples, the two effects are caused simultaneously by a common cause (e.g. the tyrant's oppression).

A related indeterminacy can arise with regard to the first type of mechanisms, those that are triggered under "generally unknown conditions." Consider again the case of an alcoholic parent. If we look at the whole population of alcoholics (or a large representative sample), suppose that their children on average drink neither more nor less than the children of non-alcoholics. Disregarding for simplicity the influence of genetic factors, this hypothetical finding might be understood in two ways. On the one hand, it could be that children of alcoholics are neither conformist nor anti-conformist: that is, their drinking behavior might be shaped by the same causes as that of children of non-alcoholics. On the other hand, it could be that half the children of alcoholics are conformist and the other half anti-conformist, leaving a net effect of zero.

Similarly, theories of voting behavior have identified both an underdog mechanism and a bandwagon mechanism. Those subject to the former tend to vote for the candidate who is behind in preelection polls, whereas those subject to the latter vote for the front-runner. If the two types are evenly mixed, there might be no noticeable net effect, so that the polls would be good predictors of the actual vote. A lack of influence of polls on voting in the aggregate would not show, however, that individuals are unaffected by the polls. Weak aggregate effects of TV violence on real-life violence could mask strong opposite effects on subgroups. In all these cases, a neutral aggregate could reflect either a homogeneous population of unaffected individuals or a heterogeneous population of individuals who are all strongly affected but in opposite directions. The need to dispel this ambiguity provides yet another argument for methodological individualism. To explain behavior at the aggregate level, we must look at the behavior of the individual components.

### Macro-mechanisms

I have been considering what we might call "atomic" mechanisms – elementary psychological reactions that cannot be reduced to other mechanisms at the same level. One might well ask how far these psychological mechanisms will take us in explaining social phenomena. The answer is that we can use atomic mechanisms as building blocks in more complex "molecular" mechanisms or *macro-mechanisms*. I offer examples throughout the book, and a general discussion in the Conclusion. Here, I give a brief preview.

Again, we may begin with proverbs. Two proverbs say, "The fear is often greater than the danger" and "Fear increases the danger." Taken together, they imply that excessive fear may create its own justification. An English proverb says that "there is a black sheep in every flock." A French proverb tells us that

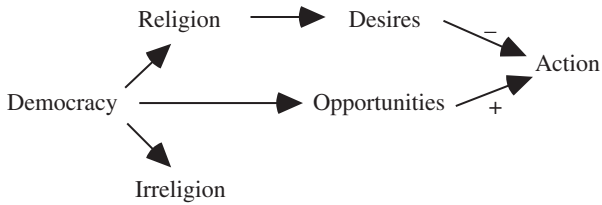



Figure 2.1

"it takes only one bad sheep to spoil a flock." Taking them together, we may infer that every flock will be spoiled.[13]

Leaving proverbs behind, let us consider another case. For centuries or millennia, elites have been wary of democracy as a regime form because they thought it would allow for all sorts of dangerous and licentious behavior. Yet opportunities for dangerous behavior will not by themselves produce such behavior: the motive must also be there. Might democratic regimes somehow restrain the desires of the citizens to do what democracy allows them to do? This was Tocqueville's claim. He thought that to satisfy a need for authority for which politics did not provide, democratic citizens would turn to religion, which tends to limit and restrain what the citizens desire.[14] The critics of democracy got it wrong, he argued, because they focused only on opportunities while neglecting desires. Although he stated this argument as if it yielded a universal law, it is more plausibly understood in terms of mechanisms. For one thing, if the spillover effect rather than the compensation effect is at work, the lack of political authority will weaken religion rather than strengthen it. For another, even if the spillover effect is at work, we cannot conclude anything about the net effect. If the opportunity set is greatly expanded and the desires only weakly restrained, the net effect of democracy may be to increase rather than reduce the incidence of the behavior in question. It is not difficult to think of examples.

The two pairs of mechanisms are summarily represented in Figure 2.1. *If* the influence of democracy on religion is mediated by the compensation effect rather than the spillover effect, democratic societies will be religious. *If* the negative effect of democracy on desires (mediated by religion) is strong

[13] I am taking a bit of a liberty with these proverbs. In its literal meaning the French phrase *une brebis galeuse* refers to a sheep with a skin disease caused by an arachnidan parasite.

[14] Tocqueville did not explain their espousal of religion by its social benefits, but, consistently with methodological individualism, by the need of *individuals* to have some authority in their lives. If the citizen "has no faith, he must serve, and if he is free, he must believe."

enough to offset the positive effect of democracy on opportunities, democratic citizens will behave moderately.[15]

### Mechanisms and laws

Often, explaining by mechanisms is the best we can do, but sometimes we can do better. Once we have identified a mechanism that is "triggered under generally unknown conditions," we may be able to identify the triggering conditions. In that case, the mechanism will be replaced by a law, albeit usually a weak one in the sense defined earlier.

Common sense assumes that a gift will make the recipient feel grateful. If he does not, we blame him. The classical moralists – from Montaigne to La Bruyère – argued that gifts tend to make recipients resentful rather than grateful. It seems that both common sense and the moralists are on to something, but they do not tell us when we can expect the one or the other outcome. A moralist from classical antiquity, Publilius Syrus, stated *triggering conditions*: a small gift creates an obligation, a large one an enemy.[16] By appealing to the size of the gift as a triggering condition, we have transformed the pair of mechanisms into a law-like statement. To cite another example, we might be able to state when a tension between a desire and a belief ("cognitive dissonance") is resolved by modifying the belief and when it is resolved by modifying the desire.[17] Purely factual beliefs may be too recalcitrant to be easily modified (Chapter 7). The person who paid $75 for a ticket to a Broadway show cannot easily fool himself into thinking he only paid $40. As noted, he may be able, however, to find some attractive aspects of the show and persuade himself that these are more important than the ones in which it is deficient.

Earlier, I mentioned the contrast between the "forbidden fruit" mechanism and the "sour grapes" mechanism. In some cases, we may be able to predict which will be triggered. In an experiment, subjects in one condition were asked to rank four records according to their attractiveness and told that the next day

[15] Other examples may be cited of peoples refraining from doing what the laws allow them to do. Montesquieu wrote that "We know that though the people of Rome assumed the right of raising plebeians to public offices, yet they never would exert this power; and though at Athens the magistrates were allowed, by the law of Aristides, to be elected from all the different classes of inhabitants, there never was a case, says Xenophon, when the common people petitioned for employments which could endanger either their security or their glory." Gibbon asserts that "the Lombards possessed freedom to elect their sovereign, and sense to decline the frequent use of that dangerous privilege." Unlike Tocqueville, neither Montesquieu nor Gibbon offers a *mechanism* to explain why these peoples pulled their punches.

[16] I am cheating a bit here, to get the example right, since Syrus refers to loans rather than to gifts. Although it seems plausible that both the loan and the gift of a large sum of money can make the recipient feel resentful, they probably do so in different ways.

[17] Recall, however, that the tension may be left unresolved.

they would receive one of them, chosen at random. Subjects in another condition ranked the records and were told that the next day they would be able to choose one of them. The next day, all subjects were told that the record they had ranked third had, for some unknown reason, become unavailable and asked to rank the four records again, as part of an attempt to discover how listening to a record for the second time might affect one's evaluation of it. As predicted by reactance theory, subjects in the first condition displayed the "sour grapes" effect by downgrading the value of the unavailable option whereas those in the second showed the "forbidden fruit" effect by upgrading it. I return to this puzzling example in Chapter 9.

Let me consider, however, a more complex example. With regard to the pair of proverbs "Absence makes the heart grow fonder" and "Out of sight, out of mind," there is actually a third proverb suggesting a triggering condition: "A short absence can do much good." La Rochefoucauld proposed a different condition: "Absence lessens moderate passions and intensifies great ones, as the wind blows out a candle but fans up a fire."[18] These plausible propositions are not very strong laws. To be able to predict the course of passion, we would have to know what counts as a short absence (three weeks?) and as a strong passion (one that keeps you awake at night?). Also, we would have to specify how duration of absence and strength of passion *interact* to generate increase or decrease of passion during an absence. Let me pursue the last issue.

### Interaction among causes

In general, the social sciences are not very good at explaining how causes interact to produce a joint effect. Most commonly, one assumes that each cause contributes separately to the effect (an "additive model"). To explain income, for instance, one may assume that it is caused in part by parental income and in part by parental education, and then use statistical methods to determine the relative contributions of these two causes. For the example I have discussed, this approach might not be adequate. The duration of the absence might not make a separate contribution to the strength of the post-absence emotion; rather its effect might depend on the strength of the pre-absence emotion. This interaction effect is shown in Figure 2.2.

Some scholars argue, though, that the world – or at least the part of it they study – simply does not exhibit many interactions of this kind. It is rarely the case, they claim, that for low levels of independent variable X the dependent

[18] To an observation by Darcy (in *Pride and Prejudice*), "I have been used to consider poetry as the *food* of love," Elizabeth Bennet responds, "Of a fine, stout, healthy love it may. Every thing nourishes what is strong already. But if it be only a slight, thin sort of inclination, I am convinced that one good sonnet will starve it entirely away."

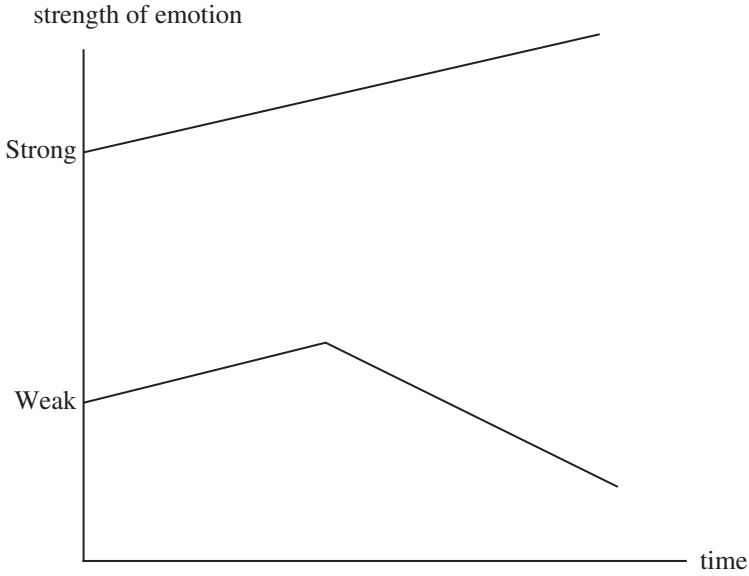



Figure 2.2

variable Z increases (decreases) with dependent variable Y, whereas for high levels of X an increase in Y causes a decrease (increase) in Z. The hypothesized relation in Figure 2.2 would (if it exists) be an exception. At most, they argue, what we find is that at low levels of X, Y has little effect on Z, whereas it does have an effect at higher levels of X. In explaining income, for instance, one may assume that parental income contributes more or less at different levels of parental education. This kind of interaction can be captured by a multiplicative interaction term so that Z is a function of X, Y, and XY. By contrast, the *reversal* of the causal effect of Y on Z at higher levels of X cannot be captured in this way. In Chapter 4 I discuss a case in which the reversal is highly plausible. Following Bentham, I argue that aptitude for public office is a multiplicative function of the official's virtue, ability, and energy. For low levels of virtue, the two other variables have a negative effect on that aptitude; for high levels, a positive effect.

The existence of an interaction effect may be subject to the same kind of indeterminacy that we find in mechanisms more generally. Consider the interaction between age and basic political attitudes as causes of extremism. One might guess that the youth organizations will be to the left of the parties themselves, giving the Young Conservatives a lighter shade of blue. Alternatively, the youth organizations of the political parties will be more extreme than the parties themselves, in which case the Young Conservatives would

have a darker shade of blue. (The young Socialists would be to the left of the party on either hypothesis.) Both guesses seem plausible, and both patterns have in fact been observed. Or consider the interaction between preconsumption mood and drug consumption as causes of postconsumption mood. One might guess that drugs such as alcohol or cocaine are *mood lifters*, attenuating depressions and turning contentment into euphoria. But one might also suspect drugs to be *mood multipliers*, making bad moods worse and good moods better. Again, both guesses seem plausible and both patterns are observed. In both cases, the first mechanism is compatible with an additive model, whereas the second implies a reversal effect.

When faced with recalcitrant data adding an interaction term, or "curve fitting," is not the only possible response. There is an alternative strategy, that of "data mining." In a curve-fitting exercise, one keeps the dependent and independent variables fixed and shops around, as it were, for a mathematical function that will give a good statistical fit. In a data-mining exercise, one keeps the mathematical function fixed (usually a simple additive model) and shops around for independent variables that have a good fit with the dependent variable. Suppose that by a "good fit" we mean a correlation that would only have a 5 percent probability of occurring by chance. In any study of a complex social phenomenon such as income, one can easily list a dozen variables that might conceivably affect it.[19] Also, there are probably half a dozen different ways of conceptualizing income. It would be very unlikely if none of the independent variables showed a correlation at the 5 percent level with one of the definitions of income.[20] The laws of probability tell us that the most improbable coincidence would be if improbable coincidences never occurred.[21]

Once a scholar has identified a suitable mathematical function or a suitable set of dependent or independent variables, she can begin to look for a causal

[19] Thus in one longitudinal study of the relation between maternal practices and child outcomes, only 35 of 552 correlations were statistically significant "at the $p < 0.05$ level" (meaning that there was one chance in twenty that they obtained by chance), a fact only evident to those who read the appendixes of the book. In the reprint edition, these were deleted.

[20] Theory may suggest that bad weather depresses stock market traders, causing them to sell. Scholars report, however, the opposite result when bad weather is defined as 100 percent cloud cover. By changing the definition of bad weather to cloud cover above 80 percent, the sign of the correlation magically reverses.

[21] I have two personal experiences. The first time I visited New York I bought tickets for two Broadway shows, one built around the music of Fats Waller and the other around that of Duke Ellington. There were few tickets left, so I had to take what I could get – which, in both shows, was row H, seat 130. This was merely uncanny, but another coincidence felt more significant. There are two experiences I have only had once. One is being invited to a dinner party and then forgetting I'd been invited. The other is being invited to a dinner party and then having the host call me up half an hour before I was supposed to be there, to tell me he had to cancel because of illness. The coincidence, which made me think for a second that someone was watching over me, is that this was one and the same party.

story to provide an intuition to back the findings. When she writes up the results for publication, the sequence is often reversed. She will state that she started with a causal theory; then looked for the most plausible way of transforming it into a formal hypothesis; and then found it confirmed by the data.[22] This is bogus science. In the natural sciences there is no need for the "logic of justification" to match or reflect "the logic of discovery." Once a hypothesis is stated in its final form, its genesis is irrelevant. What matters are its downstream consequences, not its upstream origins. This is so because the hypothesis can be tested on an indefinite number of observations over and above those that inspired the scholar to think of it in the first place. In the social sciences (and in the humanities), most explanations use a finite data set. Because procedures of data collection often are non-standardized, scholars may not be able to test their hypotheses against new data. And if procedures *are* standardized, the data may fail to reflect a changing reality. It is impossible to explain consumption patterns, for instance, without taking account of new products and of the changing prices of old ones.

There is no doubt that sharp practices of this kind occur. I do not know how common they are, only that they are sufficiently widespread to cause thoughtful social scientists to worry. The main cause of the problem is perhaps our inadequate understanding of multifactorial causality. If we had strong intuitions about how several causes can interact to produce an effect, there would be no need to rely on the mechanical procedure of "adding an interaction term" when an additive model fails. Yet because our intuitions are weak, we do not really know what to look for, and then tinkering with models seems the only alternative – at least if we retain the ambitious goal of providing law-like explanations. Given the dangers of tinkering, perhaps we should lower our ambitions instead.

### Bibliographical note

Many of the ideas in this chapter are adapted from Chapter 1 of my *Alchemies of the Mind* (Cambridge University Press, 1999). There I also cite works by Raymond Boudon, Nancy Cartwright, and Paul Veyne that advocate similar proposals. A recent statement is P. Hedström, *Dissecting the Social* (Cambridge University Press, 2005). The observation by John Paul Vann is in N. Sheehan, *A Bright Shining Lie* (New York: The Modern Library, 2009), p. 111. Useful ways of thinking about psychological mechanisms include

[22] Hence there are three problems about using correlation as a guide to causality. First, the correlation may arise purely by chance and have no causal interpretation. Second, the correlation may have an indirect causal interpretation if the two correlated phenomena are common effects of a "third factor." Third, the direction of causality might be ambiguous.

F. Heider, *The Psychology of Interpersonal Relations* (Hillsdale, NJ: Lawrence Erlbaum, 1958), and R. Abelson, *Statistics as Principled Argument* (Hillsdale, NJ: Lawrence Erlbaum, 1995). The latter also offers wise and witty remarks about the pitfalls and fallacies of statistical analysis. On these issues, two books by David Freedman are indispensable: *Statistical Models* (Cambridge University Press, 2005) and *Statistical Models and Causal Inference* (Cambridge University Press, 2009). A standard brief exposition of the idea that science explains by general laws is C. Hempel, *Philosophy of Natural Science* (Englewood Cliffs, NJ: Prentice-Hall, 1966). The principle of methodological individualism is thoroughly covered in Part 4 of M. Brodbeck (ed.), *Readings in the Philosophy of the Social Sciences* (London: Macmillan, 1969), and in Part 6 of M. Martin and L. McIntyre (eds.), *Readings in the Philosophy of Social Science* (Cambridge, MA: MIT Press, 1994); see also K. Arrow, "Methodological individualism and social knowledge," *American Economic Review: Papers and Proceedings* 84 (1994), 1–9. I have written more systematically on proverbs in "Science et sagesse: le rôle des proverbes dans la connaissance de l'homme et de la société," in J. Baechler (ed.), *L'acteur et ses raisons: Mélanges Raymond Boudon* (Paris: Presses Universitaires de France, 2000). The idea of the "psychology of tyranny" is taken from J. Roemer, "Rationalizing revolutionary ideology," *Econometrica* 53 (1985), 85–108. The study of subjects who were promised records is J. Brehm *et al.*, "The attractiveness of an eliminated choice alternative," *Journal of Experimental Social Psychology* 2 (1966), 301–13. A general introduction to reactance theory is R. Wicklund, *Freedom and Reactance* (New York: Wiley, 1974). Skepticism about interaction that induces reversal effects is found in R. Hastie and R. Dawes, *Rational Choice in an Uncertain World* (Thousand Oaks, CA: SAGE, 2001), Chapter 3. The pervasive presence of unlikely coincidences is the subject of D. Sand, *The Improbability Principle* (New York: Scientific American, 2014). The footnoted story of the 6 percent significant correlations is told in R. R. McCrae and P. T. Costa, "The paradox of parental influence," in C. Perris, W. A. Arrindell, and M. Eisemann (eds.), *Parenting and Psychopathology* (New York: Wiley), pp. 113–14. The footnoted example of the impact of bad weather on stock market traders is taken from P. Kennedy, "Oh no! I got the wrong sign! What should I do?" *Journal of Economic Education* 36 (2005), 77–92, which also contains useful comments on the costs (and benefits!) of data mining more generally.

# 3 Interpretation

## Interpretation and explanation

In many writings on the humanities, the focus has been on *interpretation* rather than explanation. In the German tradition, a contrast was often drawn between the "spiritual sciences" (*Geisteswissenschaften*) and the natural sciences (*Naturwissenschaften*). In the former, we are told, the proper procedure is that of interpretation or "understanding" (*Verstehen*). For the latter the appropriate language is that of explanation (*Erklären*). Max Weber wrote, for instance, that natural science does not aim at "understanding" the behavior of cells.

We may then ask whether the social sciences rely on understanding or on explanation. I believe this question is wrongly put. In my view, to interpret *is* to explain. Interpretation is nothing but a special case of the hypothetico-deductive method (Chapter 1). Scholars in the humanities cannot, for instance, use "empathy" as a privileged shortcut to the interpretation of behavior, since one scholar's empathetic understanding may differ from that of another. To decide among conflicting interpretations they have to confront these interpretive hunches or hypotheses (for that is what they are) with *experience*. As I argued in Chapter 1, experience includes not only the facts we are trying to understand, but also *novel facts* that we might not otherwise have thought about investigating.[1]

Interpretation is directed to human actions and to the product of human actions, such as works of art. In Chapter 16 I address the issue of interpretation of literary works, more specifically works in which we need to understand the actions of the characters as well as the choices of the author. In trying to understand other literary works, as well as the "wordless arts" of painting, sculpture, or instrumental music, this two-tier issue does not arise. Yet in these art forms too, the choices of the artist lend themselves, in principle, to much of the same analysis as that which I shall propose for authorial decisions. The

[1] In the experimental sciences, "novel facts" can mean facts that are literally new, as when one exposes rats or human beings to conditions that do not occur naturally. In the humanities and non-experimental social sciences, "novel" must be taken in the epistemic sense of "previously unsuspected" rather than in the ontological sense of "previously non-existing."

artists make choices according to some criterion of "betterness" that neither they nor we may be able to formulate explicitly, but that is revealed in practice when they discard one draft, one sketch, or one recording in favor of another. Yet the relation between the criterion of betterness and human psychology is more complicated and less well understood in the wordless arts than in (classical) fiction. I shall not attempt to deal with them.

### Rationality and intelligibility

The remainder of this chapter, therefore, will be directed to the interpretation of *action*. Interpreting an action requires us to explain it in terms of the antecedent beliefs and desires (motivations) of the agent. Moreover, we should explain these mental states themselves in a way that makes sense of them, by locating them within the full desire-belief complex. An isolated desire or belief that does not have the normal kind of solidarity with other mental states is just a brute fact that may allow us to explain behavior but not to understand it.

A paradigm mode of explaining action is to demonstrate that it was performed because it was *rational* (Chapter 13). To do so, it is not enough to show that it had good consequences for the agent: it must be understood as optimal from the agent's point of view. Trying to explain the choice by its beneficial consequences is a form of "rational-choice functionalism" – combining the two approaches I warned against in the Introduction – that sheds no light on the meaning of the behavior. It is a fact, for instance, that if people attach high value to future consequences of present behavior, that is, have a low rate of time discounting (Chapter 6), their lives go better.[2] It also seems that higher education shapes time preferences in that direction.[3] These two premises do not, however, amount to a rational-choice explanation of why people decide to become educated. For an explanation to get off the ground one would have to show that people have the requisite *beliefs* about the impact of education on the ability to delay gratification, and that they are subjectively *motivated* to acquire that ability.[4]

If behavior is rational, it is *ipso facto* also intelligible (but see Chapter 16 for an exception). Irrational behavior can also, however, be intelligible. I shall distinguish among three varieties of intelligible but irrational behavior and contrast them with some cases of unintelligible behavior.

[2] Thus five year olds who were willing to wait longer to obtain a larger marshmallow had better SAT scores when observed later.

[3] Thus when Mexican applicants to college were randomly accepted, those who were successful in the admission lottery were, when measured two years later, more patient on average.

[4] In Chapter 13 I argue that the idea of being motivated to be motivated by long-term consequences is conceptually incoherent, but that is a separate point from the one I am making here.

The first arises when the machinery of decision making (see Figure 13.1) is *truncated* in one way or another. By virtue of its peculiar urgency, a strong emotion may prevent the agent from "looking around" (i.e. gathering information) before acting. Rather than adopting a waiting strategy similar to that of the Roman general Fabius the Cunctator (hesitator), the agent rushes into action without taking the time to consider all the consequences. Another form of truncation arises in weakness of will, traditionally understood as acting against one's own better judgment (Chapter 6). The person who has decided to quit smoking yet accepts the offer of a cigarette acts on a reason, namely, a desire to smoke. For an action to be rational, however, it has to be optimal in light of the totality of reasons, not just one of them. I shall have occasion, however, to question this understanding of weakness of will.

A second variety arises in the *short-circuiting* of the machinery of decision that occurs when belief formation is biased by the agent's desires. Wishful thinking, for instance, is irrational, but fully intelligible. A subtler form of motivated belief formation arises when the agent stops gathering information when the evidence gathered so far supports the belief he would like to be true.[5] These forms of motivated belief formation are, in their way, optimizing processes: they maximize the pleasure the agent derives from his beliefs about the world rather than the pleasure he can expect from his encounters with the world.

A third variety is what we might call a *wire-crossing* in the machinery of decision. We can easily understand why the mind might engage in cognitive dissonance reduction (of which wishful thinking is one variety), but why should it also pursue dissonance *production*? The idea, cited in Chapter 2, that we believe easily what we fear is an example. Why would fear of a bad outcome make us see it as more likely than is warranted by our evidence? If the belief is supported neither by the evidence nor by our desires, why adopt it? Clearly, nothing is being optimized. In one sense such behavior is harder to understand than actions arising from truncation and short-circuiting, since *there is nothing in it* for the agent, no partial or short-term goal that it satisfies. Nevertheless it is intelligible (as I understand that idea) because it arises from the belief–desire system of the agent.

Actions that elude interpretation include those caused by compulsions and obsessions, phobic behavior, self-mutilations, anorexia, and the like. To be sure, such behavior has the effect, which explains why it is performed, of relieving the anxiety the agent feels if she does not perform it. Yet washing one's hands fifty times a day or walking up fifty flights of stairs to avoid taking the elevator is not like using a tranquilizer. Taking Valium may be as rational

[5] In an age of greater statistical innocence, Gregor Mendel, the discoverer of the laws of inheritance, apparently practiced this method of "quitting when you're ahead" in his experiments.

and intelligible as taking aspirin, but compulsive and phobic behavior is unintelligible because it is not part of an interconnected *system* of beliefs and desires. Or to take an example from John Rawls, we would find it hard to understand the behavior of someone who devoted his time to counting blades of grass unless it was linked to some other goal, such as winning a bet.

Wishful thinking is intelligible, as is counterwishful thinking. The belief of a disturbed individual that the dentist in the building next door is directing X rays at him to destroy his mind is not. By contrast, paranoid beliefs in politics are intelligible because they are rooted in the desires of the agent. A strongly anti-Semitic person is motivated to entertain absurd beliefs about the omnipotent and evil nature of the Jews (see Chapter 7). It is not that she wants Jews to *have* these features, but she is motivated to *believe* they do because the belief can rationalize her urge to destroy them. Even contradictory beliefs may be intelligible. An anti-Semite may on different occasions characterize the Jews as "vermin" *and* assert their omnipotence.[6] The very same people who say that "Jews are always trying to push in where they are not wanted" also believe that "Jews are clannish, always sticking together." One and the same Muslim may assert that the Israeli intelligence service Mossad was behind the attacks on the World Trade Center on September 11, 2001, *and* take pride in the event.

### Understanding civil wars

Let me give two extended examples to flesh out the ideas of intelligible beliefs and desires, both taken from studies of civil wars past and present. I shall then draw on the same studies and some others to address the basic hermeneutic question of how we can impute or establish motivations and beliefs.

Consider first the belief in predestination, which was a main issue dividing Calvinists and Catholics in the wars of religion. At its origin was the intense religious anxiety experienced by many believers in pre-Reformation times, due to uncertainty about their salvation. How could one be sure – could one ever be sure – that one had done enough to achieve it? Looking back to his earlier years, Calvin wrote in 1539 that even when he had satisfied the demands of the church to confess his sins and efface God's memory of them by doing good works and penance, "I was far removed from certainty and tranquility of conscience. For each time that I delved into myself or lifted my heart up to

[6] Commenting on the attitude of Virginian slaveholders, the foremost historian on the topic writes that they were against arming and freeing slaves to fight against the British in the War of 1812, arguing "that blacks were too cowardly to fight, although they also dreaded their slaves as a formidable internal enemy: living with slavery required such contradictions of belief."

You, I was struck by such an extreme horror that neither purgations nor disculpations could cure me."

What relieved him from anxiety was the shift from a conception of God as immanent in the world, an oppressive and threatening presence, to a conception of God as absolutely transcendent. Crucially, this idea was linked to the doctrine of double predestination: since God had chosen from eternity who would be saved and who would be damned, there was *nothing one could do* for one's salvation and hence no reason to worry that one had not done enough. The key interpretive issue concerns the link between this belief in predestination and the relief from anxiety. A priori, this effect of the doctrine seems unintelligible. Calvin taught that the elect were a small minority, ranging (in different statements) from one in a hundred to one in five. What could generate more anxiety than the belief that one was very likely to be among the damned and that there was nothing one could do to escape an eternity of burning in hell? Would not conversion from Catholicism to Calvinism be to go, literally, from the frying pan into the fire?

The answer is probably to be found along the lines first sketched by Max Weber. Given their belief in predestination, the Calvinists could not hold that rational and systematic effort would bring them salvation, but they could and did hold that it would give them the subjective *certainty* of salvation. Calvin himself wrote that "the vocation of the elect is like a demonstration and testimony of their election." And it seems in fact that conversion to Calvinism effectively eliminated uncertainty about salvation. I return to this form of "magical thinking" in Chapter 7. Here I merely want to emphasize how the twin mechanisms of wishful thinking and magical thinking lend intelligibility to the belief in predestination.

Consider next the intelligibility of motivations. Why have young Palestinians been willing to give their life in suicide missions? Their main motivation – to obtain or defend a national homeland – is not difficult to understand.[7] It is a cause that may be as compelling as was the defense of democracy in the struggle against Hitler. What may seem puzzling is the *strength* of the motivation. Some additional causal factors are needed to make it intelligible. I shall discuss half a dozen of these and conclude in favor of one of them.

Prior to September 11, 2001, there was a widespread belief that the typical suicide bomber in the Middle East was a single young unemployed man,

[7] By claiming that this is their main motivation, I am not denying that there may be others, such as the desire for posthumous glory or fame, the material benefits that will accrue to the family of the suicide attacker, revenge for the Israeli killing of a friend or relative, or social pressure to volunteer for a mission. As I note in the introduction to Part II, I am skeptical about the motivational power of religious benefits in the form of a privileged access to paradise.

perhaps sexually starved, for whom a religious movement could fill a vacuum that would otherwise be occupied by family and work. Then overnight, after the attack on the World Trade Center, experts on terrorism decided that they had to "rewrite the book." Even before then, however, the frequent if fluctuating deployment of female suicide bombers should have led scholars to question this stereotype. In the second Intifada, the use of female suicide bombers, some of them mothers or highly educated people, was even more striking.

The often cited factors of poverty and illiteracy also seem to have limited causal efficacy, at least as features of the individual suicide attackers. Among Palestinian suicide bombers, income and education tended in fact to be higher than in the general population. Explanations in terms of poverty are also unsatisfying because it is not clear how poverty would generate the required motivation. In one common view, the gains from blowing oneself up have to be weighed against the cost of blowing oneself up – one's life. If life is not highly valued, the cost is less. According to this approach, a life in misery and poverty is worth so little to the individual that the costs of suicide become negligible. I am skeptical about this argument, since I think that poor people find their lives as worth living as anyone else. That people adjust their aspirations to their circumstances so that they maintain a more or less constant level of satisfaction ("the hedonic treadmill") is a pretty well-established psychological finding.[8]

A more plausible factor than absolute deprivation is *relative deprivation*, that is, the gap between expectations and reality experienced by the many educated Palestinians who lacked the prospect of decent employment. Downward social mobility could have the same effect. Yet the most relevant features seem to be permanent feelings of *inferiority* and *resentment*. The first of these emotions is based on *comparison* between oneself and others, the second on *interaction* between oneself and others. Generally speaking, interaction-based emotions are more powerful than comparison-based ones. Many writers on the Palestinian suicide bombers emphasize the intense resentment caused by the daily humiliations that occur in the interaction with the Israeli forces. Beyond the degrading checks and controls to which the Palestinians are subject, there is also their awareness that many Israelis think all Arabs "lazy, cowardly, and cruel," as a Jerusalem taxi driver said to me some thirty years ago.

[8] The idea of a hedonic treadmill must be handled with some care. It must not be confused with the idea of sour grapes, that is, with the tendency of people to downgrade what they cannot have (Chapter 9). If paraplegics report being as happy after a disabling accident as before, it is surely not because they devalue the state of full mobility. Nor must it be confused with Seneca's claim that "I am not sure that [the poor] are not happier [than the rich], because they have fewer things to distract their minds," or with a similar assertion by an eighteenth-century physician, Thomas Percival, "It is one of the circumstances which soften the lot of the poor, that they are exempt from the solicitude attendant on the disposal of property."

If this account is correct, the strong resentment of those who currently occupy the desired homeland enables us to *understand* the willingness to die of the Palestinian suicide attackers. The desire to fight the Israelis derives its strength from being embedded in a larger motivational complex. There is, however, an alternative view. Palestinian suicide attackers have usually been kept on a short leash by their handlers, who are ready to provide additional pressures in case the primary motivation should fail when the time of action approaches. One would-be suicide attacker in Iraq, who was captured and disarmed because he was visibly nervous, said that for three days before his mission he had been locked up in a room with a mullah who had talked about paradise and fed him "a special soup that made him strong." The mental state that actually triggers the act of detonating the bomb may therefore be ephemeral and something of an artifact rather than a stable feature of the person. While terms such as "brainwashed" or "hypnotized" may be too strong, there is evidence that some of the attackers were in a trancelike state in the minutes before they died. When, as in such cases, an intention is isolated from the overall desire-belief system of the person, no interpretation is possible. The behavior of the handlers and more generally the organizers of the mission can, of course, be the object of interpretation.

### A hermeneutic dilemma

It is well and good to claim that behavior must be explained in terms of the antecedent mental states – desires and beliefs – that cause them, but how do we establish these prior causes? On pains of circularity, we cannot use the behavior itself as evidence. We must look at other evidence, such as statements by the agent about his motivation, the consistency of his non-verbal behavior with these statements, the motives imputed to him by others, and the consistency of *their* non-verbal behavior with these imputations. Yet how can we exclude the possibility that these verbal and non-verbal forms of behavior were purposefully chosen to make an audience believe, falsely, that a particular motivation was at work? Professions and allegations of motivations can themselves be motivated. The question is central in collective decision making. As I argue in Chapter 24, all methods for consolidating individual preferences into a social decision create incentives for the participants in the process to misrepresent their preferences.

Consider, as an example, the motives of leaders and followers in civil wars. The parties have professed, or their opponents have imputed to them, one of three motives: *religion*, *power*, and *money*. Those who profess religious motives are often accused of using them as a disguise for their real motives, be they political or pecuniary. During the French wars of religion (1562–98), the warring parties constantly accused each other of

using religion as a pretext for their political or even pecuniary aims. There were some bases for these charges. Henri de Navarre (later Henri IV) converted six times in his life, and the last conversion, in 1593, was widely suspected of opportunism. His father, Antoine de Bourbon, had already made it clear that his faith was for sale to the highest bidder. He accompanied the queen regent to mass, and his Protestant wife to communion. On his deathbed, he sought consolation from both religions. A leading reformer, Cardinal de Châtillon, married after his conversion but retained both his title as cardinal and the revenue from his bishopric. Another prelate, Antoine Carraciolo, bishop of Troyes, also wanted to combine a Protestant ministry with the income from his bishopric. A leading Catholic, Henri Duc de Guise, was perfectly willing to seek an alliance with the Calvinists against King Henri III.

In the contemporary world, too, religion is sometimes used as a pretext for politics, and politics as a pretext for money. The goals of the Chechnyan insurgents and of some Palestinian organizations, notably the Fatah, were originally exclusively political. When they took on a religious mantle, it was largely to attract a larger following. In Palestine, the rivalry with the unquestionably religious Hamas made this a necessity for organizational survival. In the Philippines, the terrorist group Abu Sayyaf has used the demand for an independent Islamic state as a pretext for kidnapping for huge ransoms. In Colombia, it remains uncertain whether the Revolutionary Armed Forces of Colombia (FARC) retains its original motivation to fight social injustice or whether it has by now degenerated into a mafia. In all these cases, as in the French wars of religion, the imputation of motives is often fraught with uncertainty. It may be hard to know, notably, whether the motives of leaders and of followers are completely concordant.

There are many reasons why people might want to misrepresent their motivations and those of their opponents. For one thing, each society has a normative hierarchy of motivations (Chapter 9) that induces a desire to present oneself as animated by a noble motivation rather than by a baser one, and to impute a low-ranked motivation to the opponent. In the French wars of religion as in the English Civil War, each side presented itself as religiously motivated and the other as merely hungry for power. For another, if one can make others attach credence to profession of a particular motivation, it may be easier to achieve one's aims. Because the image of a terrorist can be more daunting than that of a common criminal, mercenary kidnappers may increase the chances of concessions by waving the banners of a cause. In Colombia, many kidnappings are committed by common criminals who try to provoke fear among the families of victims by claiming to belong to a guerrilla group. Kidnappings are scarier if the terrorists are thought to be willing to take drastic measures if something goes wrong, and less willing to bargain over deadlines or haggle

about money. If they cannot obtain what they demand, they can at least "make a statement" by killing their victims.

The problem of self-serving bias in statements about the intentions of social agents is serious, but not insurmountable. A simple way around it might be to consider the *objective interests* of the agent and assume that in the absence of strong evidence to the contrary they coincide with her subjective motivation, regardless of what she says about the latter. Alternatively, one might identify the *actual consequences* of her action and assume that in the absence of strong evidence to the contrary they are what she intended to bring about. (Either idea might apply to the choice of higher education discussed earlier.) The fact that there exist these *two* procedures for shifting the burden of proof suggests, however, that neither is acceptable. Both objective interests and actual consequences can suggest useful hypotheses about subjective motives, but neither has a presumption in its favor.[9]

Historians and social scientists have developed other ways of handling the problem that, especially when combined, can yield reasonably certain conclusions. One technique is to go beyond statements made before an audience and to look for those less likely to be motivated by a desire for misrepresentation. Letters, diaries, reported conversations, drafts, and the like, can be invaluable sources. We know, from letters they wrote to their wives, that some delegates to the French Assemblée Constituante in 1789 voted against bicameralism and royal veto because they thought their lives might be in danger if they voted otherwise. In the assembly, they justified their votes by the public interest. During the Terror, however, the fear that private letters might be opened made them more cautious. Although contemporaries and scholars have found the motivations of Philip II of Spain impenetrable, his foremost historian writes that "one important exception exists: the dispatches of the dozen foreign ambassadors who resided at the court of Spain [and] dedicated their time and energy to removing the veil of 'secrecy and dissimulation' with which the king sought to conceal his decisions and plans from others." Hume found the report by one clergyman historian credible because "it was contrary to the interests of his order to preserve the memory" of the events he wrote about. *Lack* of interest in the outcome may also enhance credibility. Commenting on an edict by the Emperor Galerius, Gibbon wrote that "It is not usually in the language of edicts and manifestos, that we should search for the real character or the secret motives of princes, but as these were the words of a dying emperor, his situation, perhaps, may be admitted as pledge of his sincerity." In

[9] Gibbon, whom I often cite in this book, has a constant and puzzling habit of explaining behavior by a *disjunction* of motivations. To cite two examples among very many, he cites "the esteem or partiality" of a father to explain the inheritance he left to his son, and asserts that one political group was "persuaded or compelled" to acknowledge the supremacy of another. In general it is impossible to tell whether he made a backward inference from behavior to several possible motivations or whether he had some, but insufficient evidence, about each of them.

nineteenth-century England, deathbed statements were exempt from the usual rules about hearsay evidence. The first draft of a document may say more about the beliefs and motives of the author than a later published version. It is instructive, for example, to compare the drafts of Marx's *The Civil War in France* or of his letter to Vera Sassoulitch with the official versions.

There may also be a sharp contrast between what actors may say in public and what they say behind closed doors. Although the published debates of the French Assemblée Constituante in 1789–91 are endlessly fascinating, two factors conspire to make them less than reliable as evidence about mental states. On the one hand, the public setting constrained the delegates to use public-interest arguments only; naked group interest was inadmissible. On the other hand, their vanity was stimulated by speaking before a thousand fellow delegates and a thousand auditors in the galleries. In both respects, the American Federal Convention was more conducive to sincerity. Because the number of delegates was small (55, compared to 1,200 in Paris) and the proceedings were shrouded in secrecy, interest-based bargaining could and did occur. At the same time, as Madison wrote many years later, "Had the members committed themselves publicly at first, they would have afterwards supposed consistency required them to maintain their ground, whereas by secret discussion no man felt himself obliged to retain his opinions any longer than he was satisfied of their propriety and truth, and was open to the force of argument." Nor did the fear of future revelations chill the debates, as the secrecy was supposed to extend indefinitely and was in fact broken only by the publication of Madison's notes many decades later. Strategic reasons for misrepresentations are blunted if sincerity carries no cost.

In Chapter 9 I discuss how American decision makers used historical analogies to argue for or against various options in the Vietnam War. If we ask whether they relied on these analogies when making up their minds or merely used them to justify decisions reached on other grounds, a comparison between what they said in public and in private is instructive. A detailed analysis of the analogies, most of them from the crucial years of 1965 and 1966, shows some striking differences (see Table 3.1).

The comparison suggests that many of the public analogies with Cuba, Munich and Berlin were for external consumption only, while the comparisons with Korea and Malaya had greater impact on internal deliberations. (The analogy with Munich may have been downplayed in private because it was taken for granted.) The absence of public references to the perhaps most obvious comparison case, the disastrous French war in Indochina, is striking. The private references seem mainly to have been objections to George Ball's use of that analogy to argue for withdrawal from Vietnam. I also discuss the use of analogies by America decision makers in Chapter 9.

Social scientists may also remove the cost of sincerity by creating an artificial veil of ignorance. Suppose a scholar wants to study the relation

Table 3.1

| *# of analogies used in public* | | *# of analogies used in private* | |
|---|---|---|---|
| Korea | 63 | Korea | 46 |
| Munich | 42 | Dien Bien Phu | 26 |
| Greece 1947 | 32 | Malaya | 12 |
| Malaya | 22 | Munich | 10 |
| Berlin 1949, 1961 | 19 | Greece 1947 | 8 |
| Philippines | 15 | Philippines | 6 |
| Cuba | 14 | World War II | 5 |
| Turkey 1947 | 10 | Berlin 1949, 1961 | 4 |
| World War II | 9 | Cuba | 3 |
| Germany 1944 | 8 | Turkey 1947 | 2 |

between sexual orientation and some other variables of interest. It may be difficult to induce subjects to give true answers about whether they have ever had sexual experiences with members of the same sex, even if they are assured that answers will be anonymized. To get around this problem the researcher can instruct them to answer honestly in case they had any such experience and, if they never had, to flip a coin to decide whether to answer yes or no. If they comply – and they have no reason not to comply – and the sample is large enough, the data will be just as good as if everyone had answered truthfully.

Another technique to see whether the non-verbal behavior of the agents is consistent with their professed motivation is to ask: do they put their money where their mouth is? Whatever the system analysts Alain Enthoven and Daniel Ellsberg in the RAND corporation might say in public about the risks of nuclear war, their inner attitude was perhaps revealed by the fact that they decided not to sign up for RAND's lucrative retirement plan. They may have thought they would not be around long enough to benefit from it. When in 2003 the Bush administration cited its certainty that Sadaam Hussein had weapons of mass destruction as its main reason for invading Iraq, did it also deploy the requisite measures to protect American soldiers from this threat? In classical Athens, a litigator could not credibly claim to be poor if it was common knowledge that his speechwriter charged a high salary. As editor of the Soviet journal *Our Achievements*, Maxim Gorky unwittingly revealed "Our defects" in 1934, when he complained that paper shortage limited its print run. Some behavioral patterns may reveal the true motivation of kidnappers. In 1996 in Costa Rica, kidnappers (mainly, ex-contras from Nicaragua) demanded a $1 million ransom in addition to job guarantees for workers, a cut in food prices, a rise in the Costa Rican minimum wage, and the release of fellow rebels from prison. When they were offered $200,000, they were

satisfied and did not insist on the political demands, a fact that persuaded the authorities that their Robin Hood/rebel stance was a ruse and that money had always been their goal. Or take the behavior of French aristocratic émigrés in London during the French Revolution. In this hotbed of rumor about the imminent restoration of the monarchy and competition to be more-royalist-than-thou, it was vital to convey one's willingness to serve the counterrevolution. Verbal assurances were not enough. Any person who rented an apartment for more than a month was badly regarded; it was better to rent by the week to leave no doubt that one was ready to be called back to France by the counterrevolution.

Historians and social scientists routinely use such behavioral indicators to judge the sincerity of public statements. Commenting on the decision of the Emperor Diocletian to burn all books on alchemy lest the Egyptians use them to enrich themselves and rebel against the empire, Gibbon remarks that "if Diocletian had been convinced of the reality of that valuable art, far from extinguishing the memory, he would have converted the operation of it to the benefit of the public revenue." Toward the end of World War II, there was a marked degree of skepticism in occupied France about the prospects for German victory. It might not be safe to express this attitude, but it was reflected in behavior. The proportion of high school students who chose German as a foreign language (or whose parents chose it for them) doubled from 1939 to 1942 and fell rapidly thereafter. Many publishers who eagerly signed up for the right to translate German books chose not to use the option. In wartime, investors may be reluctant to state in public that they believe their country is losing, but the stock market will reveal their true beliefs.

Judges and jurors often proceed in the same way. Sometimes, they ask, "Did the accused have a motive for doing X?" hoping that an answer will help them decide whether she in fact did X. In this case, "having a motive" is an objective idea, namely, whether the accused would in some way benefit from doing X. In other cases, more relevant here, it is established that the accused did X and the question is "What was her motive for doing it?" To establish whether a killing was a crime of passion or a cold-blooded action judges and jurors do not mainly look at objective benefits, but try to establish the subjective state of mind of the accused. If the accused claims to have acted in a fit of anger or jealousy and later is shown to have bought the murder weapon ahead of time or to have taken her time over the killing,[10] her credibility is weakened.

[10] In a British decision (*R.* v. *McPherson*) from 1957, Lord Goddard asked rhetorically, "How can it be said that the appellant was acting in a gust of passion when he fired not one shot but four shots, and each shot involved the breaking of the gun to reload and the taking out of cartridges four separate times?"

Taken individually, each of these techniques may fail. A deputy might not be willing to admit to his wife that he was afraid for his life, or he might claim he was afraid in order to hide a less reputable motive (e.g. taking a bribe). In nineteenth-century India, deathbed statements were seen as unreliable since people sometimes used their dying moments to harm their enemies. In the émigré example, *both* true believers and disbelievers would be motivated to lease by the week, the former to facilitate their return to France when the day arrived and the latter to escape the criticism of being defeatist. There are limits, however, on people's ability to weave the tangled web of deceit without revealing their true motives. Hypocrisy, Somerset Maugham said, is a full-time profession. Even Tartuffe slipped in the end. To argue for the sincerity of Henri IV's religious beliefs, his biographer not only quotes the positive evidence of "numerous episodes where his religious spirit manifested itself without any advertising intention" but also argues, "Had there been any hypocrisy, it would have showed its horns on this or that pleasant occasion." Regarding Oliver Cromwell, by contrast, Hume asserted that "notwithstanding his habits of profound dissimulation, he could not so carefully guard his expressions, but that sometimes his favourite notions would escape him."

Along the same lines we may quote Montaigne:

> Those who counter what I profess by calling my frankness, my simplicity and my naturalness of manner mere artifice and cunning-prudence rather than goodness, purposive rather than natural, good sense rather than good hap – give me more honour than they take from me. They certainly make my cunning too cunning. If anyone of those men would follow me closely about and spy on me, I would declare him the winner if he does not admit that there is no teaching in his sect which would counterfeit my natural way of proceeding and keep up an appearance of such equable liberty along such tortuous paths, nor of maintaining so uncompromising a freedom of action along paths so diverse, and concede that all their striving and cleverness could never bring them to act the same.

While the benefits of misrepresentation may be considerable, the costs can be prohibitive. To some extent, the instrumental profession of motives is self-limiting. Because any given motive is embedded in a vast network of other motives and beliefs, the number of adjustments to be made in sustaining hypocrisy can be crippling. A single false note may be enough for the whole construction to crumble. Many proverbs testify to the irreversibility of the breakdown of trust. Although the folk belief "Who tells one lie will tell a hundred" needs to be severely qualified (Chapter 12), the *unqualified* belief is in fact widely held and serves to some extent as a deterrent for lying. For this reason, among others, Descartes may have been right in saying that "the greatest subtlety of all is never to make use of subtlety."

### Bibliographical note

The debate on *Erklären* versus *Verstehen* is covered in Part III of M. Martin and L. McIntyre (eds.), *Readings in the Philosophy of Social Science* (Cambridge, MA: MIT Press, 1994). The chapter in that volume by Dagfinn Føllesdal, "Hermeneutics and the hypothetico-deductive method," argues for a position close to my own. The quote from Weber is in his essay "The interpretive understanding of social action," in M. Brodbeck (ed.), *Readings in the Philosophy of the Social Sciences* (London: Macmillan, 1969), p. 33. The comment on Gregor Mendel's statistical methods is from R. Abelson, *Statistics as Principled Argument* (Hillsdale, NJ: Lawrence Erlbaum, 1995), pp. 96–7. The incoherence of anti-Semitic attitudes is touched on in J. Telushkin, *Jewish Humor* (New York: Morrow, 1992), which is also my source for other remarks on and by Jews about their alleged characteristics. The comments on the contradictory beliefs of American slaveholders is from A. Taylor, *The Internal Enemy* (New York: Norton, 2013), p. 324. The remarks on educational choice are an implicit polemic against G. Becker and C. Mulligan, "The endogenous determination of time preferences," *Quarterly Journal of Economics* 112 (1997), 729–58, further discussed in the Conclusion. The evidentiary value of deathbed confessions is discussed in J. F. Stephen, *A History of English Criminal Law* (London: Macmillan, 1883; Buffalo, NY: Hein, 1964), vol. I, pp. 447–9. H. Sass, "Affektdelikte," *Nervenarzt* 54 (1983), 557–72, lists thirteen reasons why a claim to have committed a crime out of passion might lack credibility. An outstanding interpretive discussion of motivations in the French wars of religion is D. Crouzet, *Les guerriers de Dieu* (Paris: Champ Vallon, 1990). Interpretive analyses of the motivations and beliefs of suicide attackers are found in the essays by S. Holmes, L. Ricolfi, and J. Elster in D. Gambetta (ed.), *Making Sense of Suicide Missions* (Oxford University Press, 2005). The comment on the ambassadors to Spain is in G. Parker, *The Imprudent King* (New Haven, CT: Yale University Press, 2014), p. xvi. The enumeration of public and private analogies in statements about Vietnam is a condensed version of Tables 3.1 and 3.2 in Y. Khong, *Analogies at War* (Princeton University Press, 1992). The report that Enthoven and Ellsberg did not sign up for retirement benefits is from F. Kaplan, *Wizards of Armageddon* (Stanford University Press, 1983), p. 124. The comment on *Our Achievements* is from S. Fitzpatrick, *Everyday Stalinism* (University of Chicago Press, 1999), p. 68. The behavioral indicators of the beliefs of the French in a German victory are cited from P. Burrin, *France à l'heure allemande* (Paris: Seuil, 1995). The comment on Henri IV's religious belief is in J.-P. Babelon, *Henri IV* (Paris: Fayard, 1982), p. 554. Excessive skepticism about motives is discussed in G. Mackie, "Are all men liars?" in J. Elster (ed.), *Deliberative Democracy* (Cambridge University Press, 1998).

*Part II*

# The Mind

This book is organized around the "belief–desire model" of action. To understand how people act and interact, we first have to understand how their minds work. This is largely a matter of introspection and folk psychology, refined and corrected by the more systematic studies carried out by psychologists and, increasingly, by behavioral economists. The model is vital not only for explaining behavior, but also for assigning praise, blame, or punishment. Guilt usually presupposes *mens rea*, intentions and beliefs. Strict liability – guilt assigned merely on the basis of the actual consequences of action – is rare. In fact, sometimes we hold people guilty merely on the basis of intentions even when no consequences follow. Attempted murder is a crime. "Witches," declared John Donne, "think sometimes that they kill when they do not, and are therefore as culpable as if they did." "As for witches," wrote Hobbes, "I think not that their witchcraft is any real power; but yet that they are justly punished, for the false belief they have that they can do such mischief, joined with their purpose to do it if they can."

The belief–desire model, although indispensable, is fragile. The methods we use to impute mental states to other people do not always yield stable results. If we want to measure the height of a building, it does not matter whether we do it from the roof downward or from the ground upward. In the determination of beliefs and desires, the outcome may depend on such irrelevant factors. Consider, for instance, the idea that people "maximize expected utility" (Chapter 13). To make it precise, we have to assume that they have a clear and stable idea of the *value* attached to each possible outcome of an action, and of the *probability* they assign to the occurrence of that outcome. Often that assumption is justified, but sometimes it is not.

Consider first the beliefs of the agent. In eliciting the subjective probabilities an individual attaches to an event, a standard procedure is the following. Beginning with a number p, we ask the person whether he would prefer a lottery in which he gains a certain sum of money with probability p or one in which he gains the same amount if the event in question occurs.[1] If he prefers

[1] We must assume that the event in question is one that, if it occurred, would not affect him personally, such as the discovery of life on other planets. If the event is the victory of his favorite

the former, we expose him to a new choice with the probability adjusted downward; if he prefers the latter, we adjust the probability upward. By continuing in this way, we shall ultimately reach a probability p* such that he is indifferent between a lottery in which he gains the money with probability p* and one in which he gains it if the event occurs. We can affirm, then, that the revealed or elicited probability he attaches to the event is p*. In principle, p* should be independent of the initial p: that is, the elicited probability should be independent of the procedure of elicitation. In practice, this is not the case: a higher p induces a higher p*. This finding suggests that, to some extent at least, *there is no fact of the matter*, no stable mental state that is captured by the procedure.[2]

Other procedures are even more fragile. Often, scholars impute subjective probabilities to the agents on the assumption that when they know little about the situation they will assign equal probability to each of the possible states of the world. The justification of this procedure is supposed to be the "principle of insufficient reason": if you have no positive grounds for thinking one state of the world more likely than another, logic forces you to assign equal probability to them. But states of the world can be conceptualized and counted in many ways. Suppose you are pursuing a thief and arrive at a fork in the road where three paths branch off, two going uphill and one downhill. Since you have no reason for thinking it more likely that he followed one path rather than another, the probability that he took the downhill path should, according to the principle, be one-third. But since you also have no reason for thinking he went uphill rather than downhill, the same probability should be one-half. In this case at least, the principle of insufficient reason is too indeterminate to be of any use in constructing or assigning probabilities.

Consider next the elicitation of preferences. In experiments, subjects have been asked whether they would buy various items (computer accessories, wine bottles, and the like) at a dollar figure equal to the last two digits of their social security number. Thereafter, they were asked to state the maximal price they were willing to pay for the product. It turned out that their social security number had a significant impact on what they were willing to pay. For instance, subjects who had social security numbers in the top quintile were willing to pay on average $56 for a cordless computer keyboard, while those in the bottom quintile were only willing to pay $16. Although the procedures were supposed to tap or elicit preexisting preferences, the results show that

sports team, he might bet money that it will lose so that regardless of what happens he will have something to be pleased about.

[2] This statement is probably too strong. Manipulating the procedure might elicit any probability assignment between 50 percent and 80 percent but none outside that range. In that case, we would be justified in asserting that the subject believes that the event is more likely to occur than not but is not certain that it will. This assessment is far more coarse-grained, however, than what is needed in standard models of decision making.

there was nothing there to elicit, no fact of the matter. The numbers owed more to the anchoring provided by the social security numbers than to any "real" preferences.

There is also evidence that people's *trade-offs* among values are highly unstable and may owe as much to procedural artifacts as to an underlying mental reality. Trade-offs can be captured either by *choice* or by *matching* in experiments. Subjects may be given the choice between saving many lives at a high cost per life saved (A) and saving fewer lives at lower cost (B). Alternatively, they may be asked to indicate the cost per life saved that would make them indifferent between saving the larger number of lives at that cost (option C) and option B. Suppose that a given subject states a cost lower than the cost of A. As the person is indifferent between C and B and may be assumed to prefer C to A (because C saves as many lives at lower cost), she should prefer and hence choose B over A. The overwhelming majority of subjects did in fact choose a cost for C below that of A, and yet two-thirds stated that they would choose A over B. The more important value – saving lives – is more salient in choice than in matching, although logically the two procedures should be equivalent.

There are other reasons why we should not always take statements about beliefs and other mental states at face value. Religious beliefs are especially problematic in this respect. In early seventeenth-century England, a prelate such as Bishop Andrewes could at one and the same time claim that the plague was a punishment that God imposed on sinners *and* flee London for the countryside. (By contrast, Philip II of Spain was so confident of divine support that he never made contingency plans.) The belief that by virtue of their divine origin the French kings could heal scrofula by touching the sick person was visibly withering by the end of the eighteenth century, when the traditional formula ("The king touches you; God heals you") was replaced by a subjunctive ("The king touches you; may God heal you"). The eagerness with which the king's court sought out documented proof of successful healings also suggests a belief that was not sure of itself.

For a contemporary example, consider the idea that the behavior of Islamic suicide attackers can be explained, at least in part, by their belief that there is an afterlife to which martyrdom will give them a privileged access. One may ask whether this "belief" is of the same nature as our belief that the sun will rise tomorrow, that is, whether it is used with equal confidence as a premise for action. This is not a matter of certainty versus probability, but of confidence versus lack of it. I may have great confidence in – and be willing to bet on the basis of – a probabilistic belief based on many past occurrences. The belief in the afterlife held by most people is probably not like that.[3] Rather, it may be a

[3] In one of his many ironic comments on the fragility of faith, Gibbon notes that the "bishops of Tour and Milan pronounced, without hesitation, the eternal damnation of heretics; but they were surprised, and shocked, by the bloody image of their temporal death."

somewhat shadowy "quasi-belief," held for its consumption value rather than as a premise for action. If all who claim to believe in the afterlife held the belief with full certainty, or with "confident probability," we would observe many more martyrs than we actually see. Although some believers may be of this type, and suicide attackers may be recruited disproportionately from this subset, I suspect that for many, religion serves as a consolation once the decision has been made rather than as a premise for decision.[4]

This is not to deny that faith, or superstitions, can have non-trivial effects. Thus in French Indochina, a person who thought he was pursued by a demon would run across the road just in front of a car, at the risk of being killed, because he thought the car might kill the demon who followed him in the form of his shadow. Yet note that the superstition against staying on the thirteenth floor of a hotel is not strong enough to make the hotel company incorporate an empty thirteenth floor in the building, since customers seem to be happy staying on that floor as long as it is renumbered as the fourteenth. In the Roman Empire, some individuals, including the Emperor Constantine, postponed their baptism until death was at hand, to be able to enjoy earthly pleasures while also entering heaven with a clean slate. Had they really and fully believed in the afterlife and in baptism as a necessary entry ticket, they would have taken account of the fact that death can occur suddenly without any warning signs.

Similarly, people may experience or claim to experience "quasi-emotions" that differ from genuine emotions in that they have no implications for action. Some people who claim to be indignant over third-world poverty and yet never reach for their wallet may enjoy their indignation as a consumption good, because it makes them think well of themselves. (I return to such "warm-glow" phenomena in Chapter 5.) Similarly, the visible enjoyment of many who claimed to feel grief (or "quasi-grief") after the death of Princess Diana was inconsistent with the horrible feeling of genuine grief. The appropriate term for their feelings is, I believe, "sentimentality." (The German *Schwärmerei* is even more fitting.) Oscar Wilde defined a sentimentalist as "one who desires to have the luxury of an emotion without paying for it." Whether the payment takes the form of donations to Oxfam or the form of suffering, we can tell from its absence that we are not dealing with the real thing.[5]

A related issue is the immense power of autosuggestion. Once we know that an X is supposed to be a Y, we claim *and believe* that it is obviously a Y. The

[4] The idea of religion as the "opium of the people" also suggests that it is a consumption good rather than a premise for action. It could, however, be a premise for inaction.

[5] A related phenomenon occurs when people take a third-person perspective on themselves. A proverb says, "Virtue does not know itself." Also, one cannot coherently assert one's own naivety, since the very idea presupposes lack of self-consciousness. And as Nero Wolfe put it, in one of Rex Stout's novels featuring him, "To assert dignity is to forfeit it."

world's greatest experts on Vermeer were taken in by (what *now* seem to be) obvious forgeries by van Meegeren. Proust refers to the "aptitude which enables you to discover the intentions of a symphonic piece when you have read the program, and the resemblances of a child when you know their kin." Later, I cite how Emma in Jane Austen's novel persuades herself that she is in love. A European jazz fan completely changed his high appreciation of Jack Teagarden upon learning that he was not black. If we are well disposed toward a writer, we may read deep meanings into what an impartial reader would consider trivial remarks. We project our expectations on the world and then claim that the world confirms and justifies our beliefs.

The expectations can have real effects. Thus when the same wine is served with different price labels, the pleasure centers in the brain are more highly activated in subjects who think they are drinking an expensive wine. When subjects are made to believe, falsely, that they are drinking an alcoholic beverage, they behave (up to a point) as if they were drunk. Placebo effects are the best-known examples of this phenomenon. (Even genuine painkillers are less effective when given covertly.) Placebo effects can be completely or partially reversed by opioid antagonists, suggesting that the placebo works by generating endogenous opioids in the body, similar to what produces the "runner's high." The runner, though, has to work for his pleasure.

The upshot of these remarks is that we should be wary of thinking of beliefs, desires, preferences, emotions, and the like as stable and enduring entities on a par with apples and planets. Later chapters will provide many instances of the coarse-grained, elusive, unstable, or context-dependent nature of mental states. I shall also, however, make statements that may seem to exemplify the very kind of pseudo-precision or make-believe rigor I have been warning against. In some cases, I do this to explain the internal workings of a model, without vouching for its realism. In other cases, I present a precise idea as a specification of a more general one that I believe to be valid. For example, I do not think people update their beliefs by Bayesian reasoning by the precise mechanism I describe in Chapter 13. I do believe, however, that if twelve jurors are confronted with the same evidence, those who enter the courtroom with a stronger belief in the defendant's guilt will on average be more likely to convict. Also, jurors will shift their belief toward guilt when presented with a strong piece of evidence supporting it. (By contrast, inconsistently with Bayesian reasoning they may shift their belief away from guilt when later presented by weaker evidence also supporting it.) These qualitative statements often suffice for explanation, but not for prediction. Similarly, although I do not think that people discount the future by a precise quantitative function, we can often explain their behavior by assuming that they attach more importance to immediate than to remote rewards. In some cases, we might also need to invoke the general shape of the discounting function to explain why they

sometimes change their minds, without specifying the value of the parameters. The trick – more a craft than a science – is to know how much detail to provide for a given task. When great detail is required, it is a warning sign that the task may be unfeasible.

Some final comments are called for with regard to *unconscious* mental states and mental operations. In this book I shall repeatedly refer to the unconscious workings of the mind. Dissonance reduction (Chapter 9), wishful thinking (Chapter 7), and transmutation of motives (Chapter 9), for instance, are caused by unconscious mechanisms. We may not understand well how they operate, but I find it impossible to deny that they exist. Many have also argued for the existence of unconscious mental *states*. Although these mechanisms and states cannot be accessed directly, they may, like dark matter and dark energy in the universe, be identified by their effects. In principle, one might also access them through neuroimaging.

The existence of unconscious mental mechanisms is relatively uncontroversial. The existence of unconscious mental states is a less straightforward issue. Self-deception, unlike wishful thinking, presupposes that there are unconscious beliefs. On this question, the jury is still out (see Chapter 7). Freud thought that we all have a number of unconscious and unavowable desires, as instantiated, he claimed, by the Oedipus complex. These are speculative and largely unproven ideas. The evidence for unconscious emotions and prejudices is stronger.

To the extent that unconscious mental states have causal efficacy, it should be possible to identify them by their effects. If a denial of a statement is disproportionately strong, for instance, we might infer that it is one that the person in question really, although unconsciously, believes: "The lady doth protest too much, methinks." There is a story (which I have been unable to track down) told about Sigmund Freud, who was invited to meet a prominent person, Dr. X, in the international Jewish movement. During their conversation, Dr. X asked him, "Tell me, Dr. Freud, who in your opinion is the most important Jewish personality in the world today?" Freud answered politely, "Why, I think that must be yourself, Dr. X." When Dr. X replied, "No, no," Freud asked, "Wouldn't 'No' have been enough?" Double negation can be equivalent to affirmation. The Proustian character of Legrandin (Chapter 9) offers a fine-grained illustration of this effect.

One can also identify unconscious *prejudices* by their effects. In an Implicit Association Test, experiments subjects were first asked to classify rapidly (by tapping their left or right knee) each of a list of names into those that are most often considered black (such as Malik and Lashonda) and those that are most often seen as white (such as Tiffany and Peter). Next they were asked to classify rapidly each of a list of words as pleasant in meaning (such as "love" and "baby") or unpleasant (such as "war" and "vomit"). Next, they classified a

randomly ordered list that included all of the black names, white names, pleasant words, and unpleasant words. First they were asked to tap their left knee for any black name or unpleasant-meaning word and their right knee for any white name or pleasant-meaning word. Second, the instructions were changed. They were asked to tap their left knee for white names and unpleasant words and their right knee for black names and pleasant words. It took about twice as long to respond to the second task, even though objectively the tasks were of equal difficulty. In theory, one might use such tests in a court of law, to determine whether the "disparate impact" of a racially neutral law was due to "disparate treatment" based on an unconscious prejudice.

Unconscious *emotions* can often be identified by observers who infer their existence from the characteristic physiological or behavioral expressions. Most of us have heard and many of us uttered the angry statement "I am not angry." Envy can manifest itself in a sharpness of tone and a tendency to adopt a derogatory slant that are obvious to observers but not to the subject. In *Le rouge et le noir* Mme de Rênal discovers her feelings for Julien Sorel only when she suspects that he might be in love with her chambermaid, one emotion (jealousy) thus revealing the presence of another (love).

Self-deception (see Chapter 7) is more problematic in this respect. Suppose I form and then repress the belief that my wife is having an affair with my best friend. Although unconscious, the belief that they are lovers might still guide my actions, for example, by preventing me from going to the part of town where my friend lives and where I might risk seeing my wife visiting him. This may sound like a plausible story, but to my knowledge there is no evidence that unconscious beliefs have causal efficacy. Many arguments for the existence of self-deception rely on (1) the exposure of the person to evidence strongly suggesting a belief he or she would not want to be true and (2) the fact that the agent professes and acts on a different and more palatable belief. To obtain direct evidence for the unconscious persistence of the unpalatable belief one would need to show (3) that it, too, is capable of guiding action, as in the hypothetical example just given. To repeat, I do not know of any demonstration to this effect.

My hunch is that the phenomenon does not exist. It is a pleasant conceit to imagine that my unconscious beliefs could be the handmaiden of my conscious ones, by steering me away from evidence that might undermine them, but it is no more than an unsupported just-so story. Along equally speculative lines, one could imagine that "the unconscious" is capable of inducing indirect strategies (one step backward, two steps forward), for instance, by making a child hurt herself to get the attention of her parents. These suggestions make the unconscious too similar to the conscious mind, by making it capable of having representations of the future and of other people's actions and intentions. Mental states of which we are unaware may cause spontaneous actions

such as answering, "No, no" instead of simply "No," but I do not know of any evidence that they can also cause instrumentally rational behavior.

In particular, there is no evidence that the unconscious is capable of making intertemporal trade-offs, as some economists have claimed.[6] According to one argument, workers form motivated beliefs about job safety if the benefit of holding the belief exceeds the cost. If the psychological benefit of suppressing one's fear exceeds the cost due to increased chances of accident, the worker will believe the activity to be safe. According to another argument, people may form exaggerated (but motivated) beliefs about the dangers of addiction. If I want to quit using drugs but find that my beliefs about their dangerous effects are insufficiently dissuasive, I may adopt the belief that they are more dangerous than I currently believe they are, since this belief would motivate me to suffer the withdrawal pains. Everything we know about addiction suggests the opposite, however: addicts persuade themselves that the drug is *less* dangerous than they have reason to think it is. More generally, there is no evidence that the unconscious can weigh the present benefits of false beliefs against the future costs of holding them, or the present costs of false beliefs against the future benefits of holding them.

### Bibliographical note

Evidence for anchoring of probability assessments is given in A. Tversky and D. Kahneman, "Judgment under uncertainty: heuristics and biases," *Science* 185 (1974), 1124–31. Evidence for the anchoring of preferences is given in D. Ariely, G. Loewenstein, and D. Prelec, "Coherent arbitrariness," *Quarterly Journal of Economics* 118 (2003), 73–105. The reference to Bishop Andrewes is from A. Nicolson, *God's Secretaries* (New York: HarperCollins, 2003), and that to the royal healing from M. Bloch, *Les rois thaumaturges* (Paris: Armand Colin, 1961). The examples about the relation between beliefs and behavior in Indochina and in the Roman Empire are taken from P. Veyne, *L'empire gréco-romain* (Paris: Seuil 2005), pp. 531–8. A good discussion of sentimentality is M. Tanner, "Sentimentality," *Proceedings of the Aristotelian Society* n.s. 77 (1976–7), 127–47. The evidence on the impact of price on pleasure from drinking wine is in H. Plassmann *et al.*, "Marketing actions can modulate neural representations of experienced pleasantness," *Proceedings of the National Academy of Sciences* 105 (2008), 1050–4. A large-scale (internet) experiment on unconscious prejudices is reported in B. Nosek, M. Banaji, and A. Greenwald, "Harvesting implicit group attitudes and beliefs from a demonstration website," *Group Dynamics* 6 (2002), 101–15. For the argument that

[6] As we shall see in Chapter 9, the founder of the theory of cognitive dissonance also made this claim.

Freud made the unconscious too similar to the conscious mind, see L. Naccache, *Le nouvel inconscient* (Paris: Odile Jacob, 2006). Arguments by economists about intertemporal trade-offs in the unconscious are found in G. Akerlof and W. Dickens, "The economic consequences of cognitive dissonance," *American Economic Review* 72 (1982), 307–19, and in G. Winston, "Addiction and backsliding," *Journal of Economic Behavior and Organization* 1 (1980), 295–324.

# 4 Motivations

This chapter and the two following ones will be devoted to varieties of motivation. In the present chapter, the discussion is fairly general. In the following, I focus on two specific issues, selfishness versus altruism and temporal shortsightedness versus farsightedness. These two issues complement each other to some extent, the latter being as it were the intrapersonal and intertemporal version of the former, interpersonal contrast. More importantly, they are also substantially related, in the sense that farsightedness can *mimic* altruism.

The set of human motivations is a pie that can be sliced any number of ways. Although none of them can claim canonical status, there are four approaches that I have found useful. The first proposes a continuum of motivations, the second and the third both offer a trichotomy, and the fourth a simple dichotomy. The classifications are both somewhat similar and interestingly different, allowing us to illuminate the same behavior from different angles. Following the discussion of these typologies, I offer some further comments on motivations.

### From visceral to rational

On September 11, 2001, some people jumped to their death from the World Trade Center because of the overwhelming heat. "This should not be really thought of as a choice," said Louis Garcia, New York City's chief fire marshal. "If you put people at a window and introduce that kind of heat, there's a good chance most people would feel compelled to jump." There was no real alternative. Subjectively, this may also be the experience of those who drink seawater when freshwater is unavailable. They may know that drinking even a little seawater starts you down a dangerous road: the more you drink, the thirstier you get. Yet the temptation may, for some, seem irresistible. The craving for addictive substances may also be experienced in this way. An eighteenth-century writer, Benjamin Rush, offered a dramatic illustration: "When strongly urged, by one of his friends, to leave off drinking [a habitual drunkard] said, 'Were a keg of rum in one corner of a room, and were a cannon

constantly discharging balls between me and it, I could not refrain from passing before that cannon, in order to get at the rum.'" Sexual desire may also be so overwhelming as to silence more prudential concerns.

Some emotions may also be so strong as to crowd out all other considerations. The feeling of shame, for instance, can be unbearably painful, as shown by the 1996 suicide of an American navy admiral who was about to be exposed as not entitled to some of the medals he was wearing, or by the six suicides in 1997 among Frenchmen who were exposed as consumers of pedophiliac material. Anger, too, may be overwhelmingly strong, as when Zinedine Zidane on July 9, 2006, in the last minutes of the World Cup soccer final, head-butted an Italian opponent to retaliate against a provocation, under the eyes of seventy thousand people in the stadium and an estimated one billion TV viewers worldwide. Had he paused for a fraction of a second to reflect, he would have realized that the action might cost the defeat of his team and the ruin of his reputation.

Except perhaps for the urge to jump from the World Trade Center, it is doubtful whether any of these desires was literally irresistible, in the way a boulder rolling down a hillside might be irresistible to a person trying to stop it in its course. (An urge to fall asleep may be irresistible, but falling asleep is not an action; that is why attempts to do so are self-defeating.) Addicts are somewhat sensible to costs: they consume less when prices go up.[1] People in lifeboats sometimes can prevent each other from drinking seawater. Sexual temptation and the urge to kill oneself in shame are certainly resistible. Because of their intensity, these visceral cravings nevertheless stand at one extreme of the spectrum of human motivations. They have the potential, not always realized, for blocking deliberation, trade-offs, and even choice.

At the other extreme, we have the paradigm of the rational agent who is unperturbed by visceral factors, including emotion. He acts only after having carefully – but no more carefully than is warranted under the circumstances – weighed the consequences of each available option against one another. A rational general, chief executive officer, or doctor is concerned merely with finding the best means to realize an objective goal such as winning the war, maximizing profit, or saving a life. The visceral roots of the desires do not enter into the equation.

An example of the distinction between visceral and rational motivation is provided by the difference between visceral and prudential *fear*. Although it is common to refer to fear as an emotion, it may be only a belief–desire complex. When I say, "I fear it is going to rain," I mean *only* that I believe it is going to rain and that I wish it were not going to rain. If the "fear" inspires action, as

[1] That might also be, however, because their budget does not allow them to consume at the same level (Chapter 10).

when I take an umbrella to protect me against the rain, it is a paradigm of rational behavior (Chapter 13). None of the characteristic features of the emotions (Chapter 8) is present. Visceral fear, by contrast, may induce action that is not instrumentally rational. It has been calculated, for instance, that 350 Americans who would not otherwise have died lost their lives on the road by avoiding the risk of flying after September 11, 2001. By contrast, it does not seem that the Spanish incurred excess deaths by switching from train to car after the attacks on trains in Madrid on March 11, 2004. It is possible that because of the long run of attacks by Basque Homeland and Freedom (ETA), the population had developed an attitude of prudential rather than of visceral fear toward terror bombings. For them, terrorist attacks may have been just one risk among others, similar to – albeit more dangerous than – the risk of rain.

When Franklin Roosevelt wrote that "the only thing we have to fear is . . . fear itself – nameless, unreasoning, unjustified terror which paralyzes needed efforts to convert retreat into advance," he probably had in mind rational fear of visceral fear. When Montaigne wrote "It is fear that I am most afraid of," adding that "fear banishes all wisdom from the heart," the context suggests that he referred to the same idea. In such cases, a person who rationally fears that he might at some future time be subject to irrational fear, can take precautions against the tendency, by not exposing himself to situations that might trigger fear or by preventing himself from acting out of fear. An admiral might, for instance, burn his ships to prevent himself (or his sailors) from taking flight in a panic. I discuss such strategies of "imperfect rationality" in Chapter 15.

Between the extremes of the visceral–rational continuum, we find behavior that is partly motivated by visceral factors, yet is also somewhat sensitive to cost–benefit considerations. A man may seek revenge (a visceral desire), yet also bide his time until he can catch his enemy unawares (a prudential concern). If he challenges his enemy to a duel (as required by norms of honor), he may take fencing lessons in secret (a dishonorable but useful practice). If a person is made an offer that is both unfair and advantageous, in the sense that she would be better off taking it than not, she might accept it or reject it, depending on the strength of her interest versus the strength of her resentment (Chapter 19). In more complex cases, one visceral factor might counteract another. The desire for an extramarital sexual affair might be neutralized by guilt feelings. An urge to flee from the scene of battle may be offset or preempted by an urge to fight caused by anger at the enemy, by the fear of being shamed by one's comrades, or by the fear of being shot for desertion.

### Interest, reason, and passion

In their analysis of human motivations, the seventeenth-century French moralists made a fruitful distinction among interest, reason, and passion. Interest is

the pursuit of personal advantage, be it money, fame, power, or salvation. Even action to help our children counts as the pursuit of interest, since our fate is so closely bound up with theirs. A parent who sends his children to an expensive private school where they can get the best education is not sacrificing his interest but pursuing it. The passions may be taken to include emotions as well as other visceral urges, such as hunger, thirst, and sexual or addictive cravings. The ancients also included states of madness within the same general category because, like emotions, they are involuntary, unbidden, and subversive of rational deliberation. For many purposes, we may also include states of intoxication among the passions. From the point of view of the law, anger, drunkenness, and madness have often been treated as being on a par.

Reason is a more complicated idea. The moralists mostly used it (as I shall use it here) in relation to the desire to promote the public good rather than private ends. Occasionally, they also used it to refer to long-term (prudential) motivations as distinct from short-term (myopic) concerns. Both ideas may be summarized under the heading of *impartiality*. In designing public policy, one should treat individuals impartially rather than favoring some groups or individuals over others. Individuals, too, may act on this motivation. Parents may sacrifice their interest by sending their children to a public school, because they believe in equality of opportunity. At the same time, policymakers as well as private individuals ought to treat outcomes occurring at successive times in an impartial manner by giving each of them the same weight in current decision making, rather than privileging outcomes in the near future. In fact, some moralists argued, a concern with long-term interest will also tend to promote the public good. At the Federal Convention in Philadelphia, for instance, George Mason argued that

> we ought to attend to the rights of every class of people. He had often wondered at the indifference of the superior classes of society to this dictate of humanity & policy; considering that however affluent their circumstances, or elevated their situations, might be, the course of a few years, not only might but certainly would, distribute their posterity throughout the lowest classes of Society. Every selfish motive therefore, every family attachment, ought to recommend such a system of policy as would provide no less carefully for the rights and happiness of the lowest than of the highest orders of Citizens.

Either form of impartiality has degrees. The strength of concern for others tends to vary inversely not only with genealogical distance, but with geographical remoteness. Similarly, even prudent individuals usually give somewhat more weight to the near future than to the more remote, a fact that is only partly explained by their knowledge that they might not live to enjoy the distant future.

In addition to the requirement of impartial *ends,* reason, in its consequentialist form (see later discussion), may be defined by a rational choice of *means*. Although the terms "reason" and "rationality" are sometimes used

interchangeably, I propose (Chapter 13) a conception of rationality that applies only to means, not to ends. Reason, by contrast, applies to both.

As an example of how behavior may be understood in terms of any of these three motivations, we may cite a 1783 letter from New York chancellor Robert Livingston to Alexander Hamilton in which he comments on the persecution of those who had sided with the British during the wars of independence:

> I seriously lament with you, the violent spirit of persecution which prevails here and dread its consequences upon the wealth, commerce & future tranquility of the state. I am the more hurt at it because it appears to me almost unmixed with *purer patriotic motives*. In some few it is a blind spirit of *revenge & resentment*, but in more it is the most *sordid interest*.

The phrases I have italicized correspond to reason, emotion, and interest, respectively. The adjectives are telling: reason is pure, passion is blind, interest is sordid. In Chapter 25 I illustrate the distinction with examples from constitution making.

### Id, ego, superego

In his analyses of human motivations, Freud also suggested three basic forms, each of them linked to a separate subsystem of the mind. The three systems are the id, the ego, and the superego, corresponding, respectively, to the pleasure principle, the reality principle, and conscience. The id and the superego represent, respectively, impulses and impulse control, while the ego, "helpless in both directions . . . defends itself vainly, alike against the instigations of the murderous id and against the reproaches of the punishing conscience." In a more illuminating statement from the same essay ("The Ego and the Id"), Freud wrote that the ego is "a poor creature owing service to three masters and consequently menaced by three dangers: from the external world, from the libido of the id, and from the severity of the superego." Yet even this formulation does not capture fully what I think is the useful core of Freud's idea. This is the proposition that as the ego is navigating the external world (the reality principle) it also has to fight a two-front war against the impulses from the id (pleasure principle) and the punitively severe impulse control exercised by the superego (conscience).[2]

This proposition was novel, profound, and true. What it lacks is a mechanism. Why could not the ego itself exercise whatever impulse control might be needed? Why do morality and conscience so often take the form of rigid rules? Do we need to stipulate the existence of separate and quasi-autonomous mental

[2] To combine two of Freud's metaphors, the ego is like a rider on an unruly horse (the id) who is at the same time ridden by an incubus (the superego).

functions? It took the pioneering work of George Ainslie to provide satisfactory answers to these questions. I discuss his views in Chapter 15. Here I only want to draw attention to the fact that many impulses need to be kept at bay because of the *cumulative* damage they can do if unchecked.[3] On any given occasion, drinking or eating to excess, splurging, or procrastinating (such as failing to do one's homework) need not do much harm to the agent. The damage occurs after repeated excesses (or repeated failures). The focus of impulse control, therefore, must not be the individual occasion, since the person can always say to himself or herself that a new and better life will begin tomorrow. Impulse control must address the fact that the impulse will predictably arise on an indefinite number of occasions. The solution arises from reframing the problem, so that failure to control an impulse on any one occasion is seen as a predictor of failure to control it on all later occasions. "Yes, I can postpone impulse control until tomorrow without incurring important harm or risk, but why should tomorrow be different from today? If I fail now, I shall fail tomorrow as well." By setting up an *internal domino effect* and thus raising the stakes, the agent can acquire a motivation to control her impulses that would be lacking if she just took one day at a time. The other side of the coin is that the control must be relentless and, as the Victorian moralists put it, "never suffer a single exception."

### Taking account of consequences

Finally, motivations may be consequentialist or non-consequentialist, that is, oriented either toward the outcome of action or toward the action itself. Much of economic behavior is purely consequentialist. When people put aside money for their old age or stockbrokers buy and sell shares, they attach no intrinsic value – positive or negative – to these actions themselves; they care only about the outcomes. By contrast, the unconditional pacifist who refuses to do military service even against the most evil enemy takes no account of the consequences of his behavior. What matters for him is that certain actions are unconditionally forbidden, such as taking a human life. It is not that he is *unaware* of the consequences, as may be the case in emotional action, only that consequences make no difference for what he does.

Public policy may also be adopted on either type of motivation. A policymaker might adopt the principle "Finders keepers" (e.g. in patent legislation), on the assumption that if the person who discovers a new valuable

[3] There is also a fact of cumulative *risk*. The chance of unwanted consequences from unprotected sex may be small on any given occasion, but the lifetime risk might be considerable. On any given trip, the chance of being injured in a car accident while not wearing a seatbelt is small, but the lifetime probability is about one in three.

resource is assigned the property right in it, more valuable resources will be discovered. This is a consequentialist argument. A non-consequentialist argument for the same policy might be that the person who discovers a new resource, whether it be a piece of land or a cure for cancer, has a natural *right* to property in it. For another example, we may consider the speech (XXXI) of Dion Chrysostomos against the practice of the Rhodeans to reuse old bronze statues to honor benefactors of the city: this, he argued, was both to violate the rights of those in whose honor the statues had originally been erected and to discourage potential new benefactors who knew that statues erected in their honor might soon be recycled in favor of someone else. Consequentialist arguments may (seem to) warrant harsh measures toward terrorists even if the steps that are taken violate the non-consequentialist values associated with human rights and civil liberties.[4]

A special case of non-consequentialist motivation is the principle that I shall refer to variously as everyday Kantianism, the categorical imperative, or magical thinking: *do what would be best if all did the same*. In one sense this principle is linked to consequences, since the agent does what would bring about the best outcome if everybody else did the same. These are not the consequences of *her* action, however, but of a hypothetical set of actions by her and others. In a given case, acting on the principle could have disastrous consequences for all if others do *not* follow suit. In the international arena, unilateral disarmament is an example.

Another case is the following principle of Jewish ethics. Suppose that the enemy is at the door and says, "Give me one among yourselves to be killed and we shall spare all the others; if you refuse, we shall kill you all." The Talmud requires that in such cases the Jews let themselves all be killed rather than name one to be killed so that others can be saved. If, however, the enemy says, "Give me Peter" under the same conditions, it is acceptable to hand him over. There is not a ban on causing a person to be killed to save others, but on selecting who it shall be. The novel *Sophie's Choice* presents the same dilemma.

Social norms (Chapter 21) offer a further special case of non-consequentialist behavior, with an important twist. Social norms tell people what to do, such as take revenge for an insult or refrain from eating a kid boiled in its mother's milk, not because there are any desirable results to be brought about, but because the action is mandatory in itself.[5] While not taken

[4] The parenthetical "seem to" reflects the possible operation of the "psychology of tyranny" (Chapter 2). A classical dilemma of deterrence is that the *hatred* it inspires may in the end more than offset the *fear* it is intended to cause.

[5] With regard to the rules of kosher food, of which the ban on eating a kid boiled in its mother's milk is only a historical illustration, it was thought for a while that they were justified on hygienic

to *bring about* any outcome, such actions may be seen as undertaken to *prevent* an outcome, namely, being blamed by others for not taking them. We may then ask, however, whether the blaming is also undertaken for similar consequentialist reasons. In general, I shall argue, they are not. Moreover, when people are hurt by the actions of others they retaliate even in one-shot interactions under full anonymity, such as may be obtained in experimental settings. Because the interaction is one-shot, they have nothing to gain in later encounters, and because it is anonymous, they need not fear the blame of third parties. I shall return to these experiments in several later chapters.

Even for a professed non-consequentialist, consequences may matter if they are important enough. Consider a principle that many would consider an unconditional one, the ban on torturing small children. In a "ticking bomb" scenario, imagine that a necessary and sufficient condition for preventing the detonation of a nuclear device in central Manhattan is to torture a terrorist's small child in her presence. *If* the scenario could be made credible, many non-consequentialists might acquiesce in the torture. Others would say that since the conditions in the scenario will never obtain in practice, the absolute ban remains in effect. Still others would ban the torture even if the scenario did occur. My task here is not to argue for one of these conclusions, but to make the empirical observation that in real-life situations stakes are rarely so high and knowledge rarely so certain as to force the non-consequentialist to consider the consequences of his behavior. It is possible that with more at stake or more definite knowledge he would abandon his principles, but since the situation does not arise we cannot tell for certain whether we are simply dealing with a very steep trade-off or with a total refusal to engage in trade-offs.

These four approaches to motivation capture some of the same phenomena. Visceral factors, passions, and the pleasure principle clearly have much in common. The last applies to a wider range of cases, because it involves pain avoidance as well as pleasure seeking. When students procrastinate in doing their homework, it is not necessarily because there is something else they passionately want to do. Often, they are merely taking the path of least resistance. The superego, reason, and non-consequentialist motivations also have some features in common. Although not all systems of morality are rigid and relentless, some are. Kant's moral theory is a notorious instance. (In fact, his moral philosophy may have originated in the private rules he made for himself

grounds. Today, as far as I have been able to learn, they are recommended on the grounds that it is good for one to do something that is both difficult and pointless. This idea seems to embody the fallacy of by-products that I discuss later: behavior that is justified *merely* by its character-building effects will not even have those. I do not know, though, how many of those who follow the rule do it for this reason.

to control his impulses, such as his maxim of never smoking more than one pipe after breakfast.) At the same time, morality can rise above rigidity, in individuals not subject to ambiguity aversion. The toleration of ambiguity is, in fact, often said to be the hallmark of a healthy ego. By contrast, the relation among rationality, interest, the ego, and consequentialism is more tenuous. It would be absurd to claim that the hallmark of a healthy ego is the rational pursuit of self-interest.

### Wanting and wishing

We often think of motivations as taking the form of *wanting to bring about* some state of affairs. They may also, however, take the form of *wishing some state of affairs to obtain*, whether or not there is something one can do to bring it about.[6] This distinction between wants and wishes is important if we look at the motivational component of emotion (Chapter 8). As both Seneca and Adam Smith noted, emotions can, in fact, be accompanied either by a want to do something or by a wish that something be the case. In anger or wrath, A's urge to take revenge on B cannot be satisfied by C's doing to B what A had planned to do or by B's suffering an accident.[7] What matters is not simply the outcome, that B suffer, but that he suffer by A's agency, and that B *knows* that A is the author of his suffering.[8] In the *Rhetoric*, Aristotle quotes Homer to this effect: Ulysses makes sure that Polyphemus knows who blinded him, "as if Ulysses would not have been avenged unless the Cyclops perceived both by whom and for what he had been blinded." In Racine's *Andromaque*, Hermione says that her vengeance on Pyrrhus will be "lost unless he knows, when dying, that it was I who murdered him."

By contrast, in *hatred* what matters is that the hated person or group disappear from the face of the Earth, whether this happens by my agency or by someone else's. Similarly, whereas *love* simply induces a wish for the person I love be happy, *gratitude* also makes me want the other person's happiness to be due to my agency, and for her to know that it is. In *sadism*, what matters is to *make* the other suffer; *in malice*, it is merely that the other should suffer. Adam Smith observed that a person to whom it would be

[6] Directors of funeral parlors may wish for people to die, but do not actively try to create clients. Although the wish may seem harmless, if morally repulsive, Seneca tells about a successful suit by Demades (fourth century BC) against a mortician on the grounds that he had hoped for great profits and could only have attained that aim the by the deaths of many.

[7] Hume says, though, that "the distress which Charles [VII of France] had already suffered, had tended to gratify the duke [of Burgundy]'s revenge."

[8] Experiments show, though, that A will punish B even if A knows that B will not know that he is being punished, although, perhaps surprisingly, the punishment is weaker than when A knows that B will know.

"agreeable . . . to hear that a person whom he abhorred and detested was killed by some accident . . . would reject with horror even the imagination" of causing that event. This is even clearer in *envy*. Many people who would enjoy seeing a rival's losing his possessions and would do nothing to prevent it from happening if they could, would never take active steps to destroy them, even if it could be done without costs or risks to them.[9] A person who would not set her neighbor's house on fire might abstain from calling the fire brigade if she saw it burning.

Wishful thinking (Chapter 7) is based on wishes rather than on wants. In some cases, the agent refrains from the hard work of making the world conform to his desires and adopts instead the easy path of adopting an appropriate belief about the world. If I desire to be promoted but am reluctant to make an effort, I may rely instead on insignificant signs to persuade myself that a promotion is imminent. In other cases, acting on the world is not an option. I may be unable to cause my love to be requited or my sick child to recover. In such cases, I may either engage in gratifying fantasies or face the facts. A further distinction is between cases in which the fantasies have no further consequences for action and those in which they are used as premises for behavior. I may delude myself into thinking that a woman of my acquaintance harbors a secret passion for me and yet not make any overtures to her, either because I am constrained by morality (or self-interest) or because the deluded belief is entertained mainly for its consumption value. The delusion may also be expressed overtly to its object, as happened when a secretary of John Maynard Keynes's told him that she could not help being aware of his great passion for her. Her life was ruined.

### *States that are essentially by-products*

A factor that complicates the wish-want distinction is that in some cases I can get X by doing A, but only if I do A in order to get Y. If I work hard to explain the neurophysiological basis of emotion and succeed, I may earn a high reputation. If I throw myself into work for a political cause, I may discover at the end of the process that I have also acquired a "character." If I play the piano well, I may impress others. These indirect benefits are parasitic on the main goal of the activity. If my motivation as a scholar is to earn a reputation, I am less likely to earn one. To enter a political movement *solely* for the sake of the consciousness-raising or character-building effects on oneself is doomed to fail, or will succeed only by accident. In a passage that has become proverbial, Seneca claimed that "glory follows those who avoid it." Proust noted that a

[9] Some envious people, to be sure, have no such qualms. They may live in a society where little shame attaches to envy or they may just be shameless.

musician "may sometimes betray [his true vocation] for the sake of glory, but when he seeks glory in this way, he moves further away from it, and only finds it by turning his back on it." Self-consciousness interferes with the performance. As he also wrote, although "the best way to make oneself sought after is to be hard to find," he would never give anyone advice to that effect, since "this method of achieving social success works only if one does not adopt it for that purpose."

Musical glory or social success falls in the category of *states that are essentially by-products* – states that cannot be realized by actions motivated only by the desire to realize them. These are states that may *come about*, but not be *brought about* intentionally by a simple decision. They include the desire to forget, the desire to believe, the desire to desire (such as the desire to overcome sexual impotence), the desire to sleep, the desire to laugh (one cannot tickle oneself), the desire to ignore someone,[10] and the desire to overcome stuttering. Attempts to realize these desires are likely to be ineffectual and can even make matters worse. It is a commonplace among moralists and novelists that intentional hedonism is self-defeating,[11] and that nothing engraves an experience so deeply in memory as the attempt to forget it. Although we may *wish* for these states to be realized, we should beware of *wanting* to realize them.

Many people care about *salvation* (in the afterlife) and *redemption* (for wrongs they have done). They may also believe they can achieve these goals by action. To die the death of a martyr in the fight against the infidels may provide the passport to heaven, or so some believe. To fight against the Nazis after having collaborated with them at an earlier stage may redeem the wrongdoing. Yet if these actions are undertaken for the *purpose* of achieving salvation or redemption, they may fail. In Catholic theology, the intention to buy a place in heaven by voluntary martyrdom would be an instance of the sin of simony. According to the most rigorous mystic doctrine, only work for the salvation of *others* is permitted. (Could this be a social contract: I pray for your salvation, you pray for mine?) Some Islamic scholars make a similar criticism of suicide attackers who are motivated by the belief that they will get a privileged place in paradise. Montaigne writes that when the Spartans "had to decide which of their men should individually hold the honor of having done best that day, they decided that Aristodemus had the most courageously

[10] According to La Bruyère, "A women who always has her eye on the same person, or always avoids looking at him, makes us think the same thing about her."

[11] In the final volume of *À la recherche du temps perdu* Proust, probably reflecting on his own life, wrote that the vain search for happiness can nevertheless lead to insights into the human condition that may "offer a kind of joy." The search for states that are essentially by-products can bring them about indirectly, as when a child's instructing someone to laugh causes the person to laugh out loud at this preposterous demand.

exposed himself to risk: yet they never awarded him the prize because his valor had been spurred on by his wish to purge himself of the reproach he had incurred in the battle of Thermopylae." The French press magnate Jean Prouvost, who had collaborated with the German forces during the occupation of France, tried to redeem himself by writing a large check to the resistance when it became clear that the Germans were losing the war. After Liberation, the High Court granted him a *non-lieu* (a judgment that suspends, annuls, or withdraws a case without bringing it to trial), something the Spartans presumably would not have done.[12]

The Palo Alto school of psychiatry has emphasized the importance of states that are essentially by-products, and notably how people can tie themselves into mental knots by trying to achieve directly what can only arise indirectly. Injunctions such as "Be spontaneous!" or "Don't be so obedient!" and statements such as "Nice girls don't even *think* about sex" can paralyze those to whom they are addressed. Stendhal was obsessed by the impossible goal of appearing to be *natural*. When he wrote that "all I need to be certain of achieving [social] success is to learn to show my indifference," he ignored the fact that the will to appear indifferent is inconsistent with true indifference. As Montaigne wrote, anticipating what psychologists call the "white bear effect," "there is nothing which stamps anything so vividly on our memory as the desire not to remember it."

### Push versus pull

Why do people leave one country for another? Why do academics leave one university for another? Why do peasants leave the countryside for the city? Often, answers are classified as "push versus pull." One may emigrate either because the situation at home is unbearable or because the situation abroad is irresistibly enticing – at least this is a common way of viewing the matter. In many situations, however, it is misleading. Typically, people move because they *compare* the situation at home and abroad and find that the difference is big enough to justify a move, even taking account of the costs of the move itself.[13] Yet it can make sense to distinguish push motives from pull motives, when the former are closer to the visceral end of the continuum and the latter closer to the rational end. People in the grip of strong fear sometimes run away

[12] The reason he went free was probably that the resistance needed the money and later found itself obliged to keep the tacit promise of immunity that acceptance of the check implied.

[13] This formulation presupposes that the cost of moving enters on a par with the benefits of having moved, as determinants of the overall utility of moving. Yet the costs of moving can also enter as *constraints* on the decisions. If the cheapest transatlantic fare costs more than the maximal amount a poor Italian peasant can save and borrow, he will remain in Italy no matter how much better he could do for himself in the United States (Chapter 10).

from danger rather than toward safety. The only thought in their mind is to get away, and they do not pause to think whether they might be going from the frying pan into the fire. Depending on the drug, addicts can be motivated either by the pull from euphoria (cocaine) or by the push from dysphoria (heroin). Suicidal behavior, too, may owe more to push than to pull. It is an escape from despair, not a flight to anything.

The operation of social norms (Chapter 21) can also be viewed in terms of push versus pull. The desire to excel in socially approved ways exercises a strong pull on many individuals, whether they strive for *glory* (being the best) or for *honor* (winning in a competition or combat). Other individuals are more concerned with avoiding the shame attached to the violation of social norms. In some societies, there is a general norm that says, "Don't stick your neck out." To excel in anything is to deviate, and deviation is the object of universal disapproval: "Who does he take himself for?" The relative strength of these two motivations varies across and within societies. Classical Athens illustrates the competitive striving for excellence.[14] In modern societies, small towns often show the stifling effects of the hostility to excellence. To risk a generalization, overall the push from shame seems to be a more important motivation than the pull toward glory, which is not to say that the latter cannot be powerful.

### Motivational conflict

The existence of *competing* motivations is commonplace:

I need a book so strongly that I am tempted to steal it from the library, but I also want to behave morally.

In the face of a bully I am both afraid and angry: I want to run but also to hit him.

I want all children to have public education, but I also want my child to go a private school to obtain the best education.

I want a candidate who favors legal abortion, but I also want one who favors lower taxes.

I want to smoke, but also to remain healthy.

If I am made an advantageous but unfair offer, "take it or leave it," I want both to reject it because it is unfair and to accept it because it is advantageous.

I want to donate to charity, but also to promote my own interest.

I am tempted to have an extramarital affair, but I also want to preserve my marriage.

[14] Aeschylus, for instance, wrote his plays for performance at a dramatic competition. When the young Sophocles defeated him, he was so chagrined that he left Athens for Sicily.

How is the conflict among these motivations resolved? A general answer might go as follows. Where the situation is one of "winner takes all," so that no (physical) compromise is possible, the strongest motivation wins. If my concern for my child is stronger than my concern for the schooling of children in general, I will send him or her to a private school. If my pro-choice concern is stronger than my tax-cut concern and no candidate favors both positions, I vote for a pro-choice candidate who proposes to raise taxes. If somebody offers me $3 out of a common pool of $10, intending to keep the rest for himself, I accept it. If I am offered only $2, I reject the offer. When compromise is possible, the stronger motivation has a stronger impact than the smaller one. A smoker may cut down his cigarette consumption from thirty to ten cigarettes a day. As a reflection of the strength of my altruism, I may spend 5 percent of my income on charity.[15]

This answer is not exactly wrong, but it is pretty simplistic, since the idea of "strength of motivation" is more complicated than these quick examples suggest. A motivation may owe its strength to its sheer psychic force; this is the sense in which, for instance, visceral motives are often stronger than what Madison called "the mild voice of reason." A strong motivation may also, however, be one that the agent endorses strongly because of the high value placed on it in her society. Each society or culture is in fact characterized by a normative hierarchy of motivations (Chapter 9). Other things being equal, a person would rather perform a given action for motive A than for motive B if A ranks higher in the hierarchy. These are *metamotivations*, needs to be animated by desires of a certain kind.[16] Even though weaker in the visceral sense, they may in the end win out over other motivations.

Interest and passion, notably, often show a certain *deference to reason*.[17] As Seneca said, "Reason wishes the decision that it gives to be just; anger wishes to have the decision which it has given seem the just decision." As there are very many plausible-sounding conceptions of reason, justice, and fairness, it will indeed often be possible to present a decision made in anger as conforming to reason. The trials of collaborators in countries that had been occupied by Germany during World War II were in many cases anchored in a deep desire for revenge. Yet because of their deference to reason, combined with their

[15] In Chapter 6 I discuss the more puzzling phenomenon of "loser takes all" observed in weakness of will.

[16] The idea of metamotivations is unrelated to the concept of metapreferences. An example of the latter would be a person who had two *different* preference orderings, one for eating over dieting and one for dieting over eating, and a metapreference favoring the latter. Following La Bruyère's insight that "men are very vain, and of all things hate to be thought so," a metamotivation could amount to a preference for preferring dieting to eating on grounds of health over having *the same* preference for dieting on grounds of vanity (see Chapter 9).

[17] As we shall see in Chapter 14, agents may also show a sometimes excessive deference to *rationality*.

desire to demarcate themselves from the lawless practices of the occupying regimes, the new leaders presented the severe measures as justice-based rather than emotion-based. A person may have a first-order interest in not donating to charity and a second-order desire for not seeing himself as swayed by interest only. In deference to reason, he may then adopt the philosophy of charity (Chapter 2) that can justify small donations. If others give much he will adopt a utilitarian policy that justifies small donations, and if others give little he will adopt a fairness-based policy that justifies the same behavior.

In these cases, reason has no independent causal role. It only induces an after-the-fact justification for actions already decided on other grounds. The conflict is not resolved, but swept under the carpet. I cite further examples of such rationalization in Chapter 7. In other cases, the search for a reason-based justification may change behavior. If I adopt a fairness-based policy of charity because others give little and they suddenly begin donating much more generously than before, I have to follow suit. *The same need for self-esteem* that caused me to justify self-interested behavior by impartial considerations in the first place also prevents me from changing my conception of impartiality when it no longer works in my favor. We may imagine that in *King Lear* both Burgundy and France initially fell in love with Cordelia because of her prospects, but that only the former cared so little about his self-image that he was able to shed the emotion when it no longer coincided with his interest. This is a case of interest paying deference to passion rather than to reason, suggesting that passion, or rather this particular passion, ranks above interest in the normative hierarchy. Other passions, such as envy, might well rank below interest. We might then observe efforts to undertake only such envy-based action as may be plausibly presented as interest-based. Actions that cannot be viewed in that light will not be undertaken.

Cognitive dissonance theory predicts that when one motivation is *slightly* stronger than another, it will try to recruit allies so that the reasons on one side become decisively stronger, causing any lingering doubts to disappear. The unconscious mind shops around, as it were, for additional arguments in favor of the tentative conclusion reached by the conscious mind. In such cases, "strength" of motivation cannot be taken as given, but should rather, to some extent at least, be seen as a product of the decision-making process itself. Suppose that when buying a car I attach values to differently weighted features (speed, price, comfort, appearance) of each of the alternatives and reach an overall assessment by comparing the weighted sums of the values. I might, for instance, attach overall value of 50 to brand A and of 48 to brand B. Because of the uncomfortable closeness of the comparison, I unconsciously modify the weights so that A becomes a clear winner, with 60 versus 45 as overall value. Before making the purchase, I come across brand C, which with the old weights would have scored 55, but with the new only achieves 50. Had

I seen the alternatives in the order C–A–B, I would have chosen C. Because I met them in the order A–B–C, I chose A. Such *path-dependence* undermines the simple idea that motivational conflicts are resolved according to given motivational strength.

On what I called the simplistic view, the decision whether to steal a book from the library might be represented as follows. On the one side of the balance is the benefit of being able to use the book; on the other side, the cost of guilt feelings. What I end up doing depends only on whether the cost exceeds the benefit, or vice versa. But this cannot be right, for suppose someone offered me a "guilt pill" that would remove any painful feelings of guilt for stealing the book. If guilt entered into my decisions merely as a psychic cost, it would be rational to take the pill, just as it would be rational to take a pill that would prevent me from developing a hangover from a planned drinking binge. I submit, however, that most *people would feel just as guilty about taking the pill as they would about stealing the book*.[18] I am not denying that there cannot be, in some sense, a trade-off between morality and self-interest, only that it cannot be represented in this simplistic way.

Here is a more complex case. I wish that I did not wish that I did not want to eat cream cake. I want to eat cream cake because I like it. I wish that I did not like it, because, as a moderately vain person, I think it is more important to remain slim. But I wish I were less vain. *But is that wish activated only when I want to eat cream cake*? In the conflict among my desire for cream cake, my desire to be slim, and my desire not to be vain, the first and the last can form an alliance and gang up (or sneak up) on the second. If they catch me unaware, they may succeed, but if I *understand* that the salience of my desire not to be vain is caused by the desire for cake I may be able to resist them. On another occasion, my desire for short-term gratification and my long-term desire for spontaneity may form an alliance against my medium-term desire for self-control. When more than two motives bear on the choice between two options, the idea of "strength of motivation" may be indeterminate until we know which alliance will be formed.

The seventeenth-century French moralist La Bruyère summarized two forms of motivational conflict as follows: "Nothing is easier for passion than to overcome reason; its greatest triumph is to conquer interest." We have seen that when passion "overcomes reason," it may still want to have reason on its side. Although St. Paul said, "For I do not do the good I want, but I do the evil

[18] In Chapter 13, I argue that a person who had a short time horizon would, for somewhat similar reasons, refuse to take a "discounting pill" that would make him attach more importance to future consequences of present actions. The general principle illustrated by these two pills is that a rational person would not want to do in two steps what he would not want to do in one step. He might, to be sure, want to do in two steps what he *could* not do in one.

I do not want," a more common reaction may be to persuade oneself of the goodness or justice of what, in the grip of passion, one wants to do. When passion "conquers interest," it can do so in one of two ways. The agent may, because of the *urgency* that is typical of emotion (Chapter 8), not take the time to find out where her interest lies. Alternatively, the force of emotion may be so strong that she *knowingly* acts against her interest. Such behavior may amount to weakness of will (Chapter 6).

### Material and formal preferences

Sometimes, it is useful to represent motivations as *preferences.* In Chapter 13 and in Chapter 24 I discuss how, respectively, individual and collective choice can be explained in terms of preferences. Here, I propose a distinction between material and formal preferences.

Material preferences are defined over all sorts of tangible and intangible options. In everyday life, we constantly express preferences over consumption goods, such as different varieties of fruit or different car brands, and explain our choices by citing these preferences. One may also prefer betraying one's country to betraying a friend (E. M. Forster), having loved and lost to never having loved at all (Tennyson), living in the desert to living with a contentious and angry woman (*Proverbs* 21:19), or burning out to fading away (Neil Young). Some of these preferences are wishes, defined over states of affairs, others are wants, defined over actions.

Formal preferences include *attitudes toward other people* (altruism or selfishness), *attitudes toward time* (patience or impatience), and *attitudes toward risk* (risk aversion, risk neutrality, or risk seeking).[19] I discuss them in, respectively, Chapter 5, Chapter 6, and Chapter 13. Altruism can be measured, at a first approximation, by how much personal welfare an agent is willing to sacrifice to increase the welfare of another person by one unit. Impatience can be measured by how much an agent is willing to pay for getting an immediate reward rather than suffer a delay. Suppose she is indifferent between receiving \$100 – x today or \$100 in ten days. The larger is x, the more impatient is the person. If $x = 0$, she is perfectly patient. (Are there cases in which $x < 0$?) Risk aversion can be measured by the extent an agent A has to be compensated for risk. Suppose a lottery ticket will yield either \$100 or \$0 with 50 percent probability each. An agent is willing to pay up to \$50 – x for the ticket. If $x > 0$, the agent is risk-averse, the more risk-averse the larger is x. For $x = 0$, the agent is risk-neutral. For $x < 0$, he is risk-seeking.

[19] Risk aversion should not be confused with *loss-aversion*, the tendency for losses to be weighted more heavily than equal-sized gains (Chapter 14).

A choice may involve both material and formal preferences, if I am given the choice between one orange today and two apples tomorrow or between one apple with certainty and a 50 percent chance of two oranges. We are probably less good at making such comparisons than at comparing options that differ only in the material or only in the formal respect. We are often able to state quite clearly what we prefer when other things are equal, but become confused when they are not ("trade-off aversion").

Folk psychology tends to assume that formal preferences are *character traits* that shape the behavior of an individual across the board (see Chapter 12). There is some evidence, however, that formal preferences are *domain-specific*. Most obviously, the degree of altruism depends on our closeness to the recipient. The ability to defer gratification – an expression of patience – differs when the future good is a sum of money and when it is health-related. Casual observation suggests that rock-climbers do not enjoy risk-taking in everyday life.

### Bibliographical note

A theory of visceral motivations is offered by G. Loewenstein, "Out of control: visceral influences on behavior," *Organizational Behavior and Human Decision Processes* 65 (1996), 272–92. The estimate of "excess car accidents" after September 11, 2001, is from G. Gigerenzer, "Dread risk, September 11, and fatal traffic accidents," *Psychological Science* 15 (2004), 286–7. The lack of similar excess accidents in Spain is documented in A. López-Rousseau, "Avoiding the death risk of avoiding a dread risk: the aftermath of March 11 in Spain," *Psychological Science* 16 (2005), 426–8. The trichotomy interest–reason–passion is analyzed in A. Hirschman, *The Passions and the Interests* (Princeton University Press, 1977); M. White, *Philosophy*, The Federalist, *and the Constitution* (Oxford University Press, 1987); and in my *Alchemies of the Mind* (Cambridge University Press, 1999). George Ainslie's *Picoeconomics* (Cambridge University Press, 1992) provided the lacking mechanisms for Freud's insights. A classic study of push versus pull is D. Gambetta, *Did They Jump or Were They Pushed?* (Cambridge University Press, 1983). I take the arguments of Dion Chrysostomos from P. Veyne, *L'empire gréco-romain* (Paris: Seuil, 2005), p. 217. The principle I cite from Jewish ethics is explored in D. Daube, *Collaboration with Tyranny in Rabbinic Law* (Oxford University Press, 1965), and D. Daube, *Appeasement or Resistance* (Berkeley: University of California Press, 1987). The passages from Aristotle and *Andromaque* are lifted from F. Heider, *The Psychology of Interpersonal Relations* (Hillsdale, NJ: Lawrence Erlbaum, 1958), p. 265. I develop the idea of states that are essentially by-products in Chapter 2 of *Sour Grapes* (Cambridge University Press, 1983) and apply it to the question

of redemption in "Redemption for wrongdoing," *Journal of Conflict Resolution* 50 (2006), 324–38, and to the question of salvation in "Motivations and beliefs in suicide missions," in D. Gambetta (ed.), *Making Sense of Suicide Missions* (Oxford University Press, 2005). See also L. Ross and R. Nisbett, *The Person and the Situation* (Philadelphia: Temple University Press, 1991), pp. 230–2. I discuss the "deference to reason" in trials of collaborators after World War II in Chapter 8 of *Closing the Books* (Cambridge University Press, 2004). A good introduction to the Palo Alto school of psychiatry is P. Watzlawitz, *The Pragmatics of Human Communication* (New York: Norton, 2011). Evidence for change in the weights attached to various features of alternative options is found in A. Brownstein, "Biased predecision processing," *Psychological Bulletin* 129 (2003), 545–68, and J. Brehm, "Postdecision changes in the desirability of alternatives," *Journal of Abnormal and Social Psychology* 52 (1956), 384–9. Evidence for domain-specific patience is found in G. Chapman, "Your money or your health: time preferences and trading money for health," *Medical Decision Making* 22 (2002), 410–16.

# 5 Self-interest and altruism

## Motivation and behavior

The contrast between self-interested and altruistic motivations is deceptively simple. As a first approximation, let us understand an *altruistic motivation* as the desire to enhance the welfare of others even at a net welfare loss to oneself, and an *altruistic act* as an action for which an altruistic motivation provides a sufficient reason. If I see you give money to a beggar in the street I call it an altruistic act because it is an action that *could* spring from altruistic motivations, whether or not it actually does.

For a more complex example, consider the experimental findings on "altruistic punishment." In these studies, one subject A has the option of punishing another subject B for non-cooperative behavior, at some cost to himself. There is no face-to-face interaction and the two subjects will never meet again. Yet many subjects use the punishment option, causing B to be more cooperative in his later dealings with a third party C. The punishment *could* spring from altruistic motivations, if A anticipates, and is motivated by, the benefit his punishment of B confers on C. In reality, it is more likely to be motivated by a desire for revenge.

There are many instances of such behavior outside the laboratory. In eighteenth-century France, peasants usually granted requests by beggars and vagrants for dinner and lodgings. If a peasant refused, he risked seeing his trees felled, his beasts mutilated, and his house burned down, acts of destruction that produced no benefit to the beggars and involved a risk of being caught. Although there is no reason to believe that they were in fact motivated by a desire to make the peasant take in future beggars, that motivation would be sufficient to explain them. In preindustrial England, urban food riots caused by the high prices of bread invariably ended in failure – producing nothing but "a few ruined mills and victims on the gallows," as the historian of these movements writes. Yet by virtue of their nuisance value the rebellions had a long-term success in making the propertied classes behave more moderately than they would have done otherwise.

The reason for defining altruistic motivations in terms of sacrifice of welfare rather than of material goods is to exclude cases like the following. If I pay $100,000 for my child's college education, it may be because my child's welfare is so bound up with my own that the "sacrifice" makes both of us better off.[1] The motivation, although other-regarding, is not altruistic. A case of genuine altruism would be if I sent my child to a public school when I could easily afford a private school and believed it would be better for my child. In doing so I would sacrifice not only my child's welfare but also my own. Similarly, donating to a blood bank (as distinct from giving blood to a close relative) is more likely to spring from genuinely altruistic motives. In practice, though, it may be impossible to tell whether a motivation is altruistic or merely other-regarding.

Whatever the problems of identifying altruistic motivations, there is abundant evidence of altruistic *behavior*. The Carnegie Foundation regularly hands out medals to individuals who have saved the lives of others at great risk to themselves. Many people give blood without being paid for the effort.[2] In Norway, most kidneys for transplantation are donated by relatives of the recipient. The extraction of the kidney carries a medical risk, but there is no monetary reward.[3] Many individuals, especially women, look after their old parents in addition to holding jobs and taking care of their own families. In many countries, more than half of the adult population make regular donations of money for charitable purposes. After the 2004 tsunami, high peaks of giving were observed in many developed countries. In wartime, some individuals try to disguise their disabilities or their young age so that they will be allowed to fight. The short-sighted McGeorge Bundy memorized the eye chart to get into the army. (Others mutilated themselves to get out of the army.) Many soldiers volunteer for dangerous (and some even for suicidal) missions. When people vote in national elections and thus contribute to the viability of democracy, they incur some costs and derive virtually no private benefits. The list could be extended indefinitely.

The reason why we cannot infer the existence of altruistic motivations from altruistic behavior is that other motivations may *mimic altruism*. In the terminology of Chapter 4, we may see altruism as a species of *reason*, which can be effectively simulated either by *interest* or by *passion*. (The word "mimic" or

[1] It might also be the case that I would have paid the school fees even if they were so high that the expense would make me worse off in welfare terms. In that case, payment of the lower fees is explained by other-regarding motives preempting altruism.

[2] In fact, it has been argued that non-payment is important to screen out people with infectious diseases who might donate for money.

[3] In most countries, in fact, the sale of kidneys from living persons is illegal. In this case, the motivation behind the law may be as much to protect destitute individuals against themselves as to protect recipients against low-quality body parts.

"simulate" may, but need not, imply a conscious effort to deceive others about one's real motivation.) Many people who are little concerned with being disinterested are very concerned with being praised for their disinterestedness. Thus Hume was surely wrong when he claimed that "to love the glory of virtuous deeds is a *sure* proof of the love of virtue" (my italics). A classical scholar asserts, more brutally, that among the ancient Greeks, "goodness divorced from a reputation for goodness was of limited interest." Montaigne, by contrast, asserted, "The more glittering the deed the more I subtract from its moral worth, because of the suspicion aroused in me that it was exposed more for glitter than for goodness: goods displayed are already halfway to being sold." Plutarch, Hume, Adam Smith and Schopenhauer all observed that minor actions are the most revealing of a person's character because they are less likely to be performed before an audience or, as Smith said, "less apt to be perverted by wrong systems." At the limit, the only virtuous acts are those that never come to light. The angelic grandmother of Proust's Narrator had internalized this principle so thoroughly that she attributed all her good actions to egoistic motives. To the extent that virtue has this self-effacing character, there may be more to it than meets the eye. For other reasons, to be sure, there may be less.

### Approbativeness and shamefulness

Montaigne also recognized the rarity of virtue, when he drew a distinction between true and false motivational "coins"– acting for the sake of what is right and acting for the sake of what other people think about you. As the former motivation is rare, policymakers may have to rely on the latter:

> If that false opinion [a concern for what other people think] serves the public good by keeping men to their duty . . . then let it boldly flourish and may it be fostered among us as much as it is in our power . . . Since men are not intelligent enough to be adequately paid in good coin let counterfeit coin be used as well. That method has been employed by all the lawgivers. And there is no policy which has not brought in some vain ceremonial honours, or some untruths, to keep the people to their duties.

Napoleon echoed the idea when, defending the creation of the Légion d'Honneur in 1802, he said that "by such baubles are men led." (His old soldiers from the republican army reacted strongly against this invention.) *Approbativeness* – the desire to be well thought of by others – is a false coin that may have to substitute for the true coin of altruism and morality. Alternatively, *shamefulness* – the desire not to be thought badly of by others – may serve as the false coin. Social norms may induce people to refrain from actions that they might otherwise have carried out. Abiding by the norm is not enough to make others think well of them, however. Approbation is reserved for

*supererogatory* acts, that is, those that go beyond the norm. What is obligatory in one society may be supererogatory in another. In Norway and in the United States, there is a (mild) social norm that a sibling should donate a kidney if one is needed (and suitable) for transplantation,[4] whereas in France such behavior might be seen as supererogatory. In certain social circles, donations to charity are mandatory. Thus in England in the eighteenth century, "at each level of society the principle of keeping up with the Joneses dictated the requirement to give as much and to the same causes as the Joneses," once an example had been set at the top of the social chain. Those who fell behind in their contributions were exposed in printed blacklists.

These motivations may be illustrated by two contrasting examples from eighteenth-century politics. In the first French Assemblée Constituante (1789–91), the deputies several times sacrificed important interests, ranging from giving up their feudal privileges to declaring themselves ineligible for the first ordinary legislature. Although their motivations were complex, an important component was the desire to be seen as disinterested. In the words of the biographer of one of them, they were "drunk with disinterestedness." Around the same time, in the United States, George Washington repeatedly manifested his fear that others might think he was motivated by private interest. (At the same time, he was aware that too much concern for one's virtue might appear as unvirtuous.) For another pair of illustrations, consider two conceptions of honor. According to one, honor must be *acquired*, through glorious deeds. According to another, honor is assumed as a baseline but can be *lost* through shameful deeds.

Whether approbativeness or shamefulness can mimic altruism depends on the substantive criteria others apply in assessing behavior. Some societies may place high value on – and thus stimulate the expression of – qualities that do not in any systematic way tend to mimic altruism (see Chapter 9). The desire for honor may induce all sorts of socially wasteful behavior. Napoleon's baubles were intended to encourage soldiers to risk their lives to enhance the glory of France, not to promote the welfare of the French. According to Metternich, he once said that "I grew up on the battlefield, and a man like me cares little about the lives of a million men!" Some individuals may choose a life of self-abnegation because of the praise their society bestows on religious virtuosi, but hermits and monks are often more focused on the rituals of worship than on their fellow beings. The cult of beauty in modern Western societies stimulates self-centered behavior that would seem to be inimical to the concern for others. In societies subject to what has been called "amoral

[4] In the United States, doctors often help a potential donor to resist such pressure by telling him or her early on in the process that if requested they are willing to provide a medical excuse for not donating.

familism" (southern Italy has been cited as an example) there are social norms against helping strangers in distress or against complying with the law. In *The Godfather,* Don Corleone memorably blamed his son Michael for volunteering for war service. After his son won medals for valor, "the Don … grunted disdainfully and said, 'He performs those miracles for strangers'." Overall, therefore, it is hard to say whether the desire for praise or for blame avoidance tends to mimic altruism.

### Virtue, ability, energy

One sometimes observes a tendency to confuse *virtue* and *ability*, or moral and intellectual aptitude, as Bentham called them. As I note in Chapter 22, when people are asked about their *trust* in institutions, it is rarely clear whether they assess the competence of officials or their honesty. Although many writers over the centuries have distinguished clearly between these two qualities, others have either failed to do so or assumed that ability would lead to virtue. I do not know of any evidence that the latter assumption is true, at least on the individual level. In an observation on what is known today as the ecological fallacy, Hume claimed, though, that "good morals and knowledge are almost inseparable, in every age, though not in every individual."

Bentham added a third quality to the equation: *energy*, or what he called active aptitude. He also observed, crucially, that the three qualities interact multiplicatively rather than additively to form what he called "aggregate aptitude"; hence intellectual or active aptitude are not desirable in themselves. "In so far as moral aptitude is deficient, by intellectual aptitude or active talent, both or either, in proportion to the degree in which they are present, appropriate aptitude taken in the aggregate will, instead of being increased, be diminished." He cited Napoleon as an example of a person utterly deficient in moral aptitude, but supremely endowed with intellectual and active aptitude, and Louis XVI as "one of those monarchs whose disposition was least adverse to the happiness of their fellow citizens," yet ineffective because of his "simplicity."

This conceptual scheme is simple and commonsensical, yet was not, I believe, explicitly stated before Bentham. It may have been in the air, as some quotations from his slightly older contemporaries Gibbon and Hume will show.

(i) *Virtue and ability*.[5] Commenting on Henry VIII's relations to Cardinal Wolsey, Hume wrote that "the high opinion itself, which Henry had entertained of the cardinal's capacity, tended to hasten his downfall;

[5] In a recent example of the multiplicative interaction of ability and virtue, when Roy Jenkins was asked why he was opposed to Harold Wilson's becoming leader of the British Labour Party, he "couldn't pretend that [Wilson's main rivals] show greater intellectual integrity. All Roy could say was that it was worse in Harold's case because he was more gifted."

while he imputed the bad success of that minister's undertakings, not to ill fortune or to mistake [which a less capable person might have made], but to the malignity or infidelity of his intentions." He observed that the policies of James I "were more wise and equitable, in their end, than prudent and political, in their means." By contrast, he said about Sir Henry Vane that he was "extravagant in the ends which he pursued, sagacious and profound in the means he employed."

(ii) *Ability and energy*. Hume affirmed that when the policies of Charles II took a turn for the worse, "happily, the same negligence still attended him; and, as it had lessened the influence of the good, it also diminished the affect of the bad measures, which he embraced."

(iii) *Virtue and energy*. Gibbon cited a Latin poet as comparing "in a lively epigram, the opposite characters of two præfects of Italy; he contrasts the innocent repose of a philosopher, who sometimes resigned the hours of business to slumber, perhaps to study, with the interested diligence of a rapacious minister, indefatigable in the pursuit of unjust, or sacrilegious gain. 'How happy . . . might it be for the people of Italy if Mallius could be constantly awake, and if Hadrian would always sleep!'"

Bentham may have been the first, however, to put all three qualities into the equation.[6]

Over the centuries, there have been many attempts to select voters, jurors, and deputies for their virtue and ability (more rarely for their energy). Since virtue is not an observable quality, one would need an observable proxy that can be expected to be highly correlated with it. Property, especially real estate, has been the most frequently used indicator, on the grounds that only landowners had a permanent interest in the welfare of the country. This argument was rarely made, though, by non-landowners. Ability is somewhat more amenable to direct observation and testing. Literacy, in particular, is easily ascertained, but is likely to be a necessary rather than sufficient condition for competence. In 1856 and again in 1857, two British parliamentarians made proposals to create a Public Court of Examiners to determine both the intellectual and moral qualifications for candidates in elections to the House of Commons. Unsurprisingly, they came to nothing.

## Reciprocity

Reciprocity can be a simple dyadic relation, as when each party in an ongoing relationship faces the choice between cooperating and not cooperating. One

[6] One quality is lacking in his equation: the capacity to take account of temporally remote consequences of present choices. A person may be virtuous and have great intellectual ability as well as energy, and yet her actions may amount to little if they are always oriented toward short-term ends.

farmer may harvest in August, another in September, and each can benefit from the help of the other. If the farmer whose harvest arrives first solicits the help of the other but then refuses to reciprocate in September, he is unlikely to receive assistance the following August. A stable relation of mutual assistance is likely to develop that, although it does not rely on feelings of fellowship, may foster them. During World War I, some German and British troops developed a tacit truce, a live-and-let live practice of shelling the adversary less aggressively than they could have. In this case, too, a friendly attitude toward the other side emerged over time, but as a *result* of cooperation, not as its cause.

In direct reciprocity, A helps B if and only if B has helped A. In indirect reciprocity, A helps B if B has helped C. As we shall see in later chapters, a similar distinction applies to "negative reciprocity": A may hurt B if B has hurt A, but also if B has hurt C. (Both ideas were first stated by Descartes.) The existence of indirect reciprocity suggests that people might behave altruistically in order to develop a *reputation* for having altruistic motivations. Other people will then have to decide whether the behavior reflects genuine altruism or merely a strategic desire to build a reputation for being altruistic. In this case reputation is valued on instrumental grounds, not on intrinsic ones. Whereas approbativeness causes the agent to desire esteem for its own sake, reputation is sought for the material rewards it might yield.

People may also reciprocate in one-shot situations that offer no opportunity for subsequent reward. If A behaves altruistically toward B, B may reciprocate even if both know that they will have no further interaction. The farmer harvesting in August might help the one harvesting in September even though he is planning to emigrate before the next season. One can, to be sure, imagine self-interested reasons for such reciprocation. Perhaps the farmer harvesting early fears that the other will punish him in some way if he does not reciprocate, or third parties on whose assistance he depends might ostracize him. In experimental conditions, however, one can exclude such effects. In the experimental games to be discussed later (Chapter 19), subjects interact anonymously through computer terminals, thus excluding any face-to-face effects such as shame or embarrassment. Often, the games are also designed so that a given person interacts only once with a given partner. Even under these stringent conditions, reciprocity is observed.

### Moral, social, and quasi-moral norms

I shall return to the implications of this and related experiments. Here I shall only make distinctions among three kinds of "other-regarding" motivations. *Moral norms* include the norm to help others in distress, the norm of equal sharing, and the norm of "everyday Kantianism" (do what would be best if everyone did the same). *Social norms* (Chapter 21) include norms of etiquette,

norms of revenge, and norms of queuing, drinking, and tipping. What I shall call "*quasi-moral norms*" include the norm of reciprocity (help those who help you and hurt those who hurt you) and the norm of conditional cooperation (cooperate if others do, but not otherwise). Both social norms and quasi-moral norms are conditional, in the sense that they are triggered by the presence or behavior of other people. Social norms, as I understand them, are triggered when other people can observe what the agent is doing, and quasi-moral norms when the agent can observe what other people are doing.[7] Moral norms, be they consequentialist or non-consequentialist, do not depend on either of these.

Two cases of individual responses to water shortage will illustrate the distinction between social and quasi-moral norms. In Bogotá, under the imaginative mayorship of Antanas Mockus, people followed a quasi-moral norm when reducing their consumption of water. Although individual monitoring was not feasible, the aggregate water consumption in the city was shown on TV, so that people could know whether others were for the most part complying. It appears that enough people did so to sustain the conditional cooperation. People were saying to themselves, "Since other people are cutting down on their consumption, it's only fair that I should do so as well." When there is a water shortage in California, by contrast, it seems that social norms operate to make people limit their consumption. Outdoor consumption such as watering the lawn can of course be monitored not only by neighbors, but also by municipal inspectors. Indoor consumption can be monitored by visitors, who may and do express their disapproval if the toilet bowl is clean.[8] In fact, monitoring of individual behavior also occurred in Bogotá, since children sometimes gave their parents a hard time if they did not economize on water.[9]

Quasi-moral norms can obviously be powerful in inducing altruistic behavior. Do they merely *mimic* altruism or *are* they altruistic motivations? The reason I refer to them as quasi-moral and not as moral is also why I lean to the

[7] The two can reinforce each other, when the agent can observe what the observers are themselves doing. If I see you littering, I may not mind your watching me doing the same. If I see you carefully putting your ice cream wrapper in your pocket, however, fairness and fear of disapproval may combine to produce conformity (see also Chapter 21).

[8] Saving water is also a concern in normal times. In New York City, it is achieved by laws fixing the maximal volume of toilet cisterns. In some countries, water consumption is monitored and taxed by the public authorities. In much of Europe, it is established by having toilets with two push buttons dispensing different amounts of water for different uses. The latter system is interesting in that it operates neither on opportunities nor on incentives (Chapter 10), only on the unobservable goodwill of the person.

[9] Experimental findings also suggest this mechanism. In an energy-saving campaign, signs were posted in shower rooms urging students to save energy by turning their shower off as they soaped themselves, and turning it on only to rinse themselves. The signs had minimal effect. When one or two experimental confederates started complying, however, compliance by other shower users increased dramatically. Although the confederates did not say anything to the others, their behavior might serve as a tacit reproach to non-compliers.

first answer. The norm of reciprocity allows you *not* to help others in distress unless they have helped you previously. A typical moral norm is to help others in distress unconditionally, even if there is no prior history of assistance. The norm of conditional cooperation allows one to use normal amounts of water if nobody else is reducing consumption, whereas both utilitarianism and everyday Kantianism would endorse unilateral reduction. Moral norms, one might say, are *proactive*; quasi-moral norms, only *reactive*. Another way of expressing the difference is that the feeling of injustice seems to have stronger motivational force than the sense of justice. As we shall see later (Chapter 19), proposals that Responders in an experiment tend to reject as unfair, with the consequence that neither they nor the Proposers get anything, are of the same order of magnitude as what Proposers tend to offer when unconstrained by the fear of rejection.

It would seem that we could identify the operation of genuinely altruistic motives if two conditions are satisfied. First, the action benefiting others is proactive, not reactive. Second, it is anonymous, in the sense that the identity of the benevolent actor is known neither to the beneficiary nor to third parties.[10] We may imagine, for instance, a person sending an anonymous money order to the charity Oxfam or dropping money into the collection box of an empty church. The second example is not as clear-cut as one would want, since the person might be motivated by his belief that God observes him and will reward him. The belief may be illogical (an instance of the "by-product fallacy") but might still be quite common. The first example might seem more unambiguous. Yet even the purest acts of altruism such as anonymous donations to strangers may stem from murky motives. According to Kant,

> it is absolutely impossible to make out by experience with complete certainty a single case in which the maxim of an action, however right in itself, rested simply on moral grounds and on the conception of duty. Sometimes it happens that with the sharpest self-examination we can find nothing beside the moral principle of duty which could have been powerful enough to move us to this or that action and to so great a sacrifice; yet we cannot from this infer with certainty that it was not really some secret impulse of self-love, under the false appearance of duty, that was the actual determining cause of the will. We like to flatter ourselves by falsely taking credit for a more noble motive; whereas in fact we can never, even by the strictest examination, get completely behind the secret springs of action; since, when the question is of moral worth, it is not with the actions which we see that we are concerned, but with those inward principles of them which we do not.

Kant is saying that even if we are not performing before an external audience, we can never know whether we are playing to the *internal audience*, to use a

[10] In experiments, the identity of the subject is hidden to the experimenter. In donations to charity, it is hidden to the officials in the charitable organization.

metaphor whose meaning will become clearer, I hope, by the examples I give in Chapter 9.[11] The act of hiding one's virtue to others that Montaigne found so virtuous will be apparent to oneself. As La Rochefoucauld noted, amour-propre "always finds compensations, and even when it gives up vanity it loses nothing." As he also said, "If pure love exists, free from the dross of our other passions, it lies hidden in the depths of our hearts and unknown even to ourselves." At best, said Proust, we may be able to learn our true motives from others: "We are familiar only with the passions of others, and what we come to know about our own, we have been able to learn only from them. Upon ourselves, they act only indirectly, by way of our imagination, which substitutes for our primary motives alternative motives that are more acceptable." I return to this important idea of a substitution, or transmutation, in Chapter 9.

Both Kant and Proust were influenced by the French moralists of the seventeenth century, notably by La Rochefoucauld. We can state their framework as follows. All individuals some of the time and some individuals all the time are *egoistic*, motivated only by their private material benefits. Many more individuals are *egocentric*, motivated by material self-interest, but also by two forms of amour-propre: vanity and the desire for self-approval. People want be approved by an external audience, even, as Seneca noted, by those of whom they disapprove. Amour-propre can also make people seek the approval of the internal audience. In either case, the approval may require some sacrifice of material self-interest. Yet the egocentric does not really care about others: they matter only as spectators of, or conditions for, his sacrifice. Some people, finally, are genuinely *altruistic*, and as such indifferent to both the external and the internal audience. The Narrator in Proust's *Recherche* offers a description, when he notes that:

> when, in the course of my life, I have had occasion to meet with, in convents for instance, literally saintly examples of practical charity, they have generally had the brisk, decided, undisturbed, and slightly brutal air of a busy surgeon, the face in which one can discern no commiseration, no tenderness at the sight of suffering humanity, and no fear of hurting it, which is the face devoid of gentleness, the antipathetic and sublime face of true goodness.

### The warm glow (the Valmont effect)

I pursue this issue by means of an example. When people donate to good causes, what motivates them? For specificity, consider donations to the charity

[11] Hume asserts that when George Fox, the founder of the Quakers, "had been sufficiently consecrated in his own imagination, he felt that the fumes of self-applause soon dissipate, if not continually supplied by the admiration of others." I do not know of any studies that have examined this question in a more general perspective.

organization Save the Children. If donors are motivated exclusively by altruism, that is, by their desire to increase the welfare of children, they face a collective action problem (Chapter 23). For altruistic donors, the welfare of the recipients is in fact a public good, on a par with a clean environment. What matters for each potential donor is that children be better off, not that *he* make them better off (Chapter 4). He benefits as much from the donations of others as from his own, just as an environmentalist benefits as much from the cleaning-up efforts of others as from his own.

In terms to be explained later (Chapter 18), economists assume that individual donations have to form a *Nash equilibrium*. Each altruist donates the amount that is optimal, from his perspective, given how much others give. In deciding how much to donate, he does not take account, however, of the benefits he provides to other potential donors. If philanthropic donations were constrained to form an equilibrium in this sense, predicted levels of charitable giving would be much lower than what we actually observe. Moreover, private donations should be crowded out by government interventions, also contrary to what we observe. Economists have responded to these problems by assuming that donations are motivated by a *private good,* the "warm glow" from giving.

An early reference to the warm-glow effect occurs in the sulfurous work by Choderlos de Laclos, *Les liaisons dangereuses.* In one scene, the cynical rake Vicomte de Valmont engages in charitable behavior for the purely selfish motive of seducing the Présidente de Tourvel. Knowing that one of her servants is following him to observe his behavior, he seeks out a poor family whose property is about to be taken from them to pay for tax arrears:

> I summon the tax collector. And, giving in to my generous compassion, I nobly part with fifty-six livres, for which paltry sum five human beings were being reduced to straw and poverty. After this simple little action you may imagine what a chorus of blessings echoed all around me … In the midst of the blessings from this family, I looked not unlike a hero in the final act of a drama. You will not forget that my faithful spy was there in the crowd. My aim was accomplished … After all, I am very pleased with my idea. I have no doubt that this woman is worth making a great effort for. One day this will count for something in her eyes.

In his further reflection on his experience, Valmont also discovers the intrinsic pleasure of doing good. When the family members kneel before him to express their thanks, he discovers a strange sensation: "I shall confess to a momentary weakness. My eyes filled with tears and I felt within me an involuntary but delightful emotion. I am astonished at the pleasure one feels at doing good. And I should be tempted to believe that those whom we call virtuous do not have so much merit as we are led to believe." Thus Valmont *explains* the seemingly virtuous actions of others – but not his own – by the warm-glow effect.

The explanation raises empirical as well as conceptual questions. Empirically, the cited objection to an altruistic equilibrium assumes that each potential donor is *aware* of the amount other people donate. This assumption seems unrealistic. Moreover, altruism and the warm glow do not exhaust the set of possible motivations for charitable behavior. After a natural disaster, a tsunami or an earthquake, people donate *more* when they learn from the media that others are donating heavily, consistently with the operation of a quasi-moral norm. In politics, too, this mechanism can be at work. Thus in a 1963 fundraising campaign for Barry Goldwater, the newspaper "*Roll Call* reported that the Goldwater campaign already had $7.5 million in the bank. It was really $125,000, although [the campaign director] hardly minded the publicity; it would only bring in more."

The conceptual questions are more serious. Let me try to reconstruct the mindset of a typical – not, I hope, caricatural – economist. Faced with any observed behavior, she will first try to explain it by assuming that the agent was motivated by rational self-interest. In cases such as donations to charity or voting in national elections, this line of argument is clearly unpromising. Generally speaking, the natural reaction to an explanatory failure is to try to explain the observed facts by departing as little as possible from the original model. In the case that concerns me here, the smallest deviation from *rational egoism* might seem to be that of *rational egocentricity.* The agent might be concerned both with her own material benefit and the degree to which she can think of herself as a moral person. As in other cases, there would be a trade-off between these aims: as La Rochefoucauld suggested, she might be willing to sacrifice some material welfare to get the warm glow from the enhanced self-image.

Warm-glow theorists of philanthropy, voting, and similar non-selfish actions do not seem to realize, however, that this small adjustment to the model, *substituting egocentricity for egoism*, requires another and more radical one: *substituting irrationality for rationality*. Specifically, agents have to deceive themselves about their own motives. This substitution is required by what I take to be a conceptual truth: one cannot derive a warm glow from an action unless the agent believes that the action was performed at least in part to benefit others. *An egocentric agent who performs for the inner audience has to believe that she is altruistic.* An agent who performed "good actions" only for the *conscious end* of enhancing her self-image could not achieve that aim, any more than one can enhance one's self-image by paying another person to praise oneself.[12]

[12] In a phrase from Beaumarchais's play *Le mariage de Figaro* that is now the motto of the newspaper *Le Figaro*, "sans la liberté de blâmer il n'est pas d'éloge flatteur" (without the freedom to criticize, there is no true praise). Two years before the play opened, Gibbon wrote

Moreover, introspection and casual observation suggest that the strength of the warm glow depends on *how much* we donate to others. I cannot easily persuade myself that I have an altruistic character merely by giving a quarter to the panhandler in the street. The greater the benefit to others (and the greater the cost to oneself), the warmer the glow. In practice, and perhaps in principle, it would be hard to distinguish between the enhanced welfare of others as the *altruistic goal* of the donor and its role as a *condition for achieving his egocentric goal*.

### Imputing motivations

In addition to the agent's own motivation, the explanation of her behavior must often appeal to her beliefs about the motivations of others. In forming these beliefs she faces the same hermeneutic dilemma as does the historian or the social scientist. Since she cannot take the professed motivations of others at face value, she can use triangulations of the general kind I discussed in Chapter 3. In addition, she can deploy techniques that only apply to face-to-face interactions. Other people may be able to identify an amateur liar by his body language (or lack of it), since concentration on what he is saying causes him to neglect the gestures that normally accompany spontaneous speech. Also, to verify professed motives one may set a trap for the agent. Whereas historians are unable to trap the individuals they are studying and social scientists are usually prevented on ethical grounds from doing so, an employer, a spouse, or a parent may feel less constrained.

The imputation of motives to others is often tainted by malice. Given the choice between believing that an altruistic action was caused by an altruistic motivation and that it was based on self-interest, we often assume the latter even if there are no positive grounds for the belief. Although such distrust can make sense for prudential reasons, in many cases this justification is unavailable. Gossip, for instance, seems often to be motivated by what the French moralists, following Augustine, called the *malignity* and *weakness* of human nature.[13]

that "the slaves who would not dare censure [Emperor Julian's] defects, were not worthy to applaud his virtues." Although this conceptual truth has not prevented dictators from ordering underlings and newspaper editors to praise them, its violation points to their fundamental irrationality. Montaigne cites Carneades as saying that "that the only thing which the sons of princes really learned properly was horsemanship, since in all other sports men yield to them and allow them to win whereas a horse is neither a flatterer nor a courtier." Tiberius, who according to Tacitus detested flattery and criticism in equal measure, was an exception. When he asked the advice of the Senate, it very wisely refused to give it: agreeing and disagreeing would have been equally disastrous.

[13] I disagree with those who want to *explain* gossip by its role in enforcing social norms. True, gossip can act as a multiplier on the informal sanctions that sustain social norms, but I believe its origin is more deep-seated.

According to La Rochefoucauld, "If we had no faults we should not find so much enjoyment in seeing faults in others." In fact, as he also wrote, our desire to find faults in others is so strong that it often helps us to find them: "Our enemies are nearer the truth in their opinion of us than we are ourselves." Yet, even if our enemies are closer to the truth, they err, too, even if they err less, from the opposite direction. On a scale from 0 to 10, if I am 6, I will think I am 9, and my enemies will think I am 4.

For an analysis of this attitude – sometimes called the "hermeneutics of suspicion" – I can do no better than quote from Jeremy Bentham (translated from his clumsy French):

> Whatever position the King [Louis XVI] takes, whatever sacrifices he makes, he will never succeed in silencing these slanderers: they are a vermin that bad temper and vanity will never fail to nourish in even the most healthy political body. It is first and foremost vanity that is the most prolific source of this injustice. One wants to deal subtly with everything . . . and prefers the most contrived assumption to the shame of having suspected that the behavior of a public person might have a laudable motive. If Washington persists in his retirement, it can only be a means to use the road through anarchy to open up the path to despotism. If Necker instead of accepting payment for his services like anyone else pays with his own funds for being allowed to render them, it can only be a sophisticated means to satisfy his greed. If Louis XVI abdicates the legislative power in favor of his people, it can only be as the result of an elaborate plan to take it all back and even more in a favorable moment.

An irony is that the last of the specious accusations cited in this text (written in early 1789) was probably justified by the fall of 1790. One of the king's closest advisers, Saint-Priest, wrote that by that time he had stopped resisting the encroachments of the legislature because "he had convinced himself that the Assembly would be discredited through its own errors." Conspiracy theories can be accurate, because conspiracies exist. Yet the tendency to find them may owe less to experience than to a malignant reluctance to admit that public figures might act for good reasons.

In Chapter 16 I consider other examples of the hermeneutics of suspicion, related to the interpretation of texts.

### Bibliographical note

This chapter draws on my "Altruistic motivations and altruistic behavior," in S. C. Kolm and J. M. Ythier (eds.), *Handbook on the Economics of Giving, Reciprocity and Altruism*, vol. I (Amsterdam: Elsevier, 2006). Other chapters in this volume, notably the introductory essay by Kolm, provide a wealth of empirical information and theoretical analysis; the essays in vol. II are equally informative. The reference to French beggars is from G. Lefebvre, *La grande peur* (Paris: Armand Colin, 1988), p. 40, and that to English peasant rebellions

from E. P. Thompson, "The moral economy of the English crowd in the 18th century," *Past and Present* 80 (1971), 76–136. On attitudes toward kidney donation, see H. Lorenzen and F. Paterson, "Donations from the living: are the French and Norwegians altruistic?" in J. Elster and N. Herpin (eds.), *The Ethics of Medical Choice* (London: Pinter, 1994). I take the idea (and the word) of "approbativeness" from A. O. Lovejoy, *Reflections on Human Nature* (Baltimore: Johns Hopkins University Press, 1961). The role of disinterestedness in the French Revolution is discussed in B. M. Shapiro, "Self-sacrifice, self-interest, or self-defense? The constituent assembly and the 'self-denying ordinance' of May 1791," *French Historical Studies* 25 (2002), 625–56. For the American parallel, see G. Wood, "Interest and disinterestedness in the making of the constitution," in R. Beeman, S. Botein, and E. Carter II (eds.), *Beyond Confederation: Origins of the Constitution and American National Identity* (Chapel Hill: University of North Carolina Press, 1987). The comment on mandatory giving to charity in eighteenth-century England is from P. Langford, *Public Life and the Propertied Englishman* (Oxford University Press, 1991), pp. 564–5. The anecdote about Roy Jenkins is from J. Campbell, *Roy Jenkins: A Rounded Life* (London: Jonathan Cape, 2014), p. 228. For the proposal to require moral and intellectual tests of English MPs, see H. Witmer, *The Property Qualifications of Members of Parliament* (New York: Columbia University Press, 1943), p. 201. The tit-for-tat example from World War I is taken from R. Axelrod, *The Evolution of Cooperation* (New York: Basic Books, 1984). For the Trust Game, see C. Camerer, *Behavioral Game Theory* (New York: Russell Sage, 2004), Chapter 2.7. An extensive discussion of warm-glow giving is in J. Andreoni, "Philanthropy," in S.-C. Kolm and J. M. Ythier (eds.), *Handbook of the Economics of Giving, Altruism and Reciprocity*, vol. II (Amsterdam: North-Holland, 2006), pp. 1202–69. I criticize the warm-glow account in "The Valmont effect," in P. Illingworth, T. Pogge, and L. Wenar (eds.), *Giving Well: The Ethics of Philanthropy* (Oxford University Press, 2011), 67–83. A warm-glow theory of voting is offered in B. Caplan, *The Myth of the Rational Voter* (Princeton University Press, 2007); for a criticism, see J. Elster and H. Landemore, "Ideology and dystopia," *Critical Review* 20 (2008), 273–89. The reference to the Goldwater campaign is in R. Perlstein, *Before the Storm* (New York: Nation Books, 2001), p. 320. On detecting lies, see P. Ekman, *Telling Lies* (New York: Norton, 1992). The passage from Bentham is taken from his *Rights, Representation, and Reform* (Oxford University Press, 2002), pp. 17–18.

# 6 Myopia and foresight

In the previous chapter I explored the relation between self and others. In the present chapter, the focus is on the relation between the present and future. In some cases, and to some extent, the analogy between *other selves* and *future states of the self* can be useful.[1] We may, that is, look for intrapersonal and intertemporal analogies to interpersonal relations. The questions "If not me, who?" and "If not now, when?" have a common root in magical thinking (Chapter 14). Drawing on the same analogy, La Bruyère observed that "To think only of oneself and of the present time is a source of error in politics." The economic idea of externalities has a parallel in the idea of internalities (Chapter 17). Economists sometimes treat prudence and altruism within the same conceptual framework. The problem of time inconsistency can arise in intrapersonal as well as interpersonal contexts (see Chapter 18). Although the analogy is endlessly fascinating, it has to be handled with care. It may suggest hypotheses, but does not lend them any support.

### Beyond gradient climbing

Freud's pleasure principle (Chapter 4) is the tendency to seek immediate gratification of desires. One manifestation of this tendency is the adoption of the belief one would like to be true rather than the belief that is supported by the evidence. Wishful thinking makes me feel good here and now, even if it may cause me to fall flat on my face later on. Another manifestation occurs in the choice between two actions that induce different temporal utility streams. The pleasure principle dictates the choice of the stream that has the highest utility in the first period, regardless of the shape of the streams in later periods.

More generally, a decision maker, be it an earthworm or a firm, may engage in *gradient climbing*. At any point in time it scans the *nearby* options to see whether one of them yields greater *immediate* benefits than the status quo. The restriction to nearby options is a form of "spatial myopia": out of sight, out of

[1] The phrase "future selves" is acceptable only as a metaphor. The phrase "multiple selves" does not even have that value (in non-pathological cases).

mind. The restriction to immediate benefits is a form of temporal myopia: the pleasure principle. The earthworm scans the environment to see whether any spot nearby is more humid than the one it is currently occupying and moves to that spot if it finds one. The firm scans the "space" of routines that are close to what it is currently doing to find one that promises better short-term performance and adopts it if it finds one. After a while, the earthworm or the firm may come to rest in a place that is superior (in the short run) to all nearby positions. It has attained a *local maximum*.

Human beings can do better. Intentionality – the ability to re-present the absent – enables us to go beyond the pleasure principle and take account of the temporally remote consequences of present choices. Planning ahead enables us to make choices that have better consequences than those that would flow from minute-by-minute or second-by-second decisions. In some cases, such far-sighted actions may be undertaken to satisfy current needs better, as when an alcoholic forgoes having a drink in a nearby restaurant so that he can buy a whole bottle in a remotely located store at the same price. In other cases, the actions are undertaken to satisfy future needs, as when I save for my old age. Whereas the former kind of foresight is also observed in non-human animals, the latter has usually been thought to be beyond their capacity. Some evidence suggests, however, that primates may be able to plan on the basis of expected rather than actual needs. Be that as it may, acting on the basis of projected needs is obviously a more sophisticated operation.

Let me give four examples of acting on the basis of temporally remote consequences. The first three examples are also discussed in later chapters.

***Reculer pour mieux sauter***. This French phrase, the rough equivalent of "one step backward, two steps forward," is illustrated by the fundamental fact of economic life that to invest for greater consumption in the future one must consume less in the present. The agent accepts a state that is inferior to the status quo because it is a condition for realizing a superior alternative later on. Needless to say, this makes sense only if (1) the inferior state allows the agent to survive and (2) the present value of the gains from the superior state are large enough to justify the loss involved in moving to the inferior state.

**Waiting**. Many wines, although good from the time they are bottled, improve with age. To benefit from this fact, the agent has to be willing to reject an option (drinking the wine right away) that is superior to the status quo because the rejection is a condition for realizing an even better outcome later. Again, deferring consumption might not always make sense, for instance, if the agent does not expect to live long enough to enjoy the improved wine. For a more consequential example, consider the choice of spouse. Rather than proposing marriage or accepting a marriage proposal on the first occasion an acceptable candidate appears, one might wait for somebody even better suited.

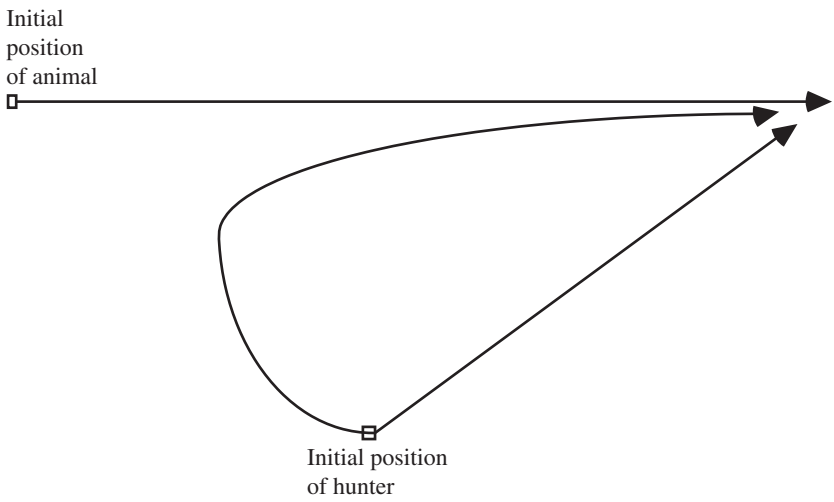



Figure 6.1

The risk, abundantly illustrated in world literature, is that nobody better suited might come along.

**Shooting ahead of the target**. To hit a moving target, one should not aim at where it is, but at where it will be at the time of encounter. Similarly, to pursue a moving target, one should aim in a straight line at where the target will be rather than follow the curved path induced by always aiming at its current position.

In Figure 6.1, the hunter, even if he is moving somewhat more slowly than the animal, can catch up with it by going in a straight line toward the point where it will be at some calculable time in the future. If, however, he always aims in the direction of the current position of the animal, following the curved path in the diagram, he will never catch up with it. As we shall see (Chapter 11), natural selection in a changing environment can be viewed in this perspective.

**A straight line is not always the fastest way**. When trying to reach a stationary target, a straight line is not always the most efficient path. In Figure 6.2, the rescuer might impulsively run straight toward the drowning swimmer until she reaches the shoreline and then swim the remaining distance. If she had paused (but not too long!) to reflect, however, she might have realized that as she can run faster than she can swim, she would reach the swimmer faster by taking an indirect path that, although longer on the whole, has a shorter stretch in the water. We behave in this way when we take a turnpike rather than the road that, on the map, seems to be the shorter. In economic planning such "turnpike behavior" is often optimal.

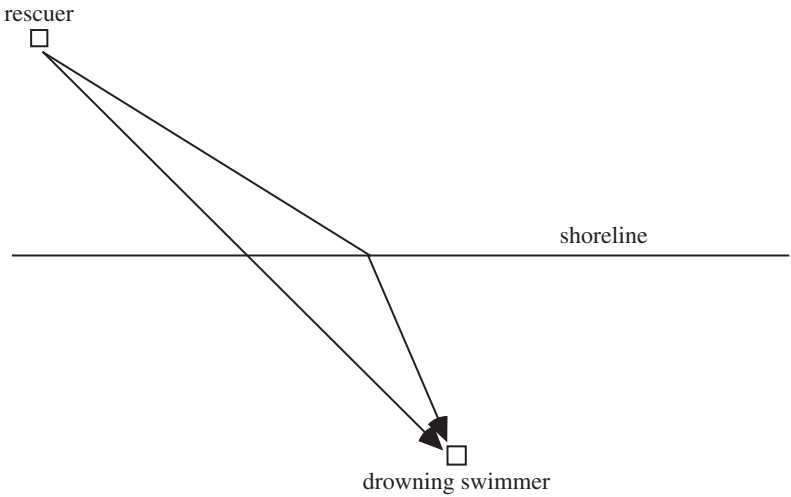



Figure 6.2

### Time discounting

The existence of the capacity for long-term planning does not imply that it will be used. For perceived long-term consequences to make a difference for present behavior, agents must be *motivated* to take them into account. In the language of psychologists, they must be ready to *defer gratification*. In the language of economists, they must not be subject to excessive *time discounting*.[2] The cognitive and motivational elements are both needed. If future outcomes are shrouded in uncertainty, they cannot motivate present behavior. If they involve risk, their motivating force is also attenuated. The ability of future outcomes to shape present behavior is affected both by *the time at which* and by *the probability with which* they will occur. The formal preferences by which they affect choice are, respectively, time discounting and risk attitudes.

As the phrase suggests, time discounting (or myopia) is the tendency to attach less importance to rewards in the distant future than to rewards in the near future or in the present.[3] The tendency is pretty universal: "If a hangover

[2] In this book, the phrase "a high rate of time discounting" shall mean that future rewards have a small present value. The phrase "a high discount factor" shall mean that they have a large present value. To illustrate and motivate this seemingly strange terminology, assume that the agent is indifferent between three units of reward tomorrow and two units today. The future reward is discounted (reduced) by one-third. The discount factor (the number by which we have to multiply the future reward to get its present value) is two-thirds.

[3] Some individuals, such as pathological misers, may attach more importance to future than to present utility. For them, the time to consume is never quite ripe.

came before we got drunk we would see that we never drank to excess" (Montaigne). If given the choice between $100 today and $110 a year from today, most people would probably, even in the absence of inflation, prefer the former. This preference could, however, have a number of sources.

*Profit-seeking.* Some people might prefer the early reward because they can invest the funds and withdraw more than $110 in a year's time.

*Scarcity.* Others might take the $100 now because they need the money to survive. Getting a bigger sum later has no value if they expect to be dead by then. Or suppose I have the choice between catching fish in the stream with my hands and making a net that will enable me to catch many more fish. Because I cannot catch fish while making the net, however, the opportunity cost of making the net may be so high that I cannot afford it.

*Mortality.* Still others might take the smaller reward because they have a disease that entails a 10 percent chance of dying within a year. More generally, when planning for the future we have to take account of the fact that we know that we shall die but not when.

*Risk aversion.* If the future sum is an expected reward, involving a 50 percent chance of $130 and a 50 percent chance of $90, risk aversion might induce a preference for getting $100 with certainty today.

*Pure time discounting.* Finally, some people might prefer the early reward simply because it arrives earlier. Just as a big house seen in the distance appears to be smaller than a small house close up, a large sum in the future may appear, subjectively, as smaller than a small sum in the present. In the following, I shall consider only this case.

Is pure time discounting irrational? Suppose a person discounts future rewards very heavily. Rather than getting a college education, which involves a temporary sacrifice of income with a higher income later on, he takes a low-level job with few promotion possibilities immediately after high school. Because he ignores the long-term impact of smoking and of tasty but unhealthy food, he has a short life expectancy.[4] If he does not respect the law on moral grounds, prudential considerations will not deter him from violating it. It is quite likely, in other words, that his life will be short and miserable. If this is not irrational behavior, what is?

In my view, pure time discounting, by itself, is not irrational. It may cause the agent's life to go worse than if she cared more about the future, but that may also be true of selfish motivations. Someone who only cares for herself may end up having a sad and impoverished life. As Montaigne said, "He who

[4] Fifty years ago many people might have "ignored" these consequences in the sense of being unaware of them. While this is less likely today, they may still "ignore" them in the sense of attaching less importance to them in their decisions. Not infrequently, they may also be in a state of "motivated ignorance" (a form of wishful thinking) about the consequences.

does not live a little for others hardly lives at all for himself." We should not for that reason, however, say that selfishness is irrational. I pursue these questions in Chapter 13. Here, I focus on the proper way to conceptualize time discounting. Several approaches, which have radically different implications, are available.

To model time discounting, decision theorists traditionally assumed that people discount future utility *exponentially*. One unit of utility t periods in the future has a present value of $k^t$, where $k < 1$ is the per-period discount factor. Exponential discounting has the attractive factor, from a normative point of view, that it allows *consistent planning*. If one stream of rewards has a greater present value than another at one point in time, it will have a greater present value at all other points in time. Hence the agent is never subject to a preference reversal, which is usually (in the absence of reasons for changing one's mind) taken as a hallmark of irrationality.

Empirically, however, the notion of consistent planning makes less sense. Casual observation shows, and systematic observation confirms, that most of us are frequently subject to preference reversal. We often fail to carry out intentions to save, do exercises in the morning, do our piano practice, keep our appointments, and so on. I may call my dentist on March 1 to make an appointment for April 1, only to call and cancel on March 30, saying (untruthfully) that I have to go to a funeral. To account for these varieties of everyday irrationality (and for a large number of other phenomena) we can replace the assumption of exponential discounting with that of *hyperbolic discounting*.

Suppose that the discounted present value of 1 unit of utility t periods into the future equals $1/(1 + kt)$. (In the example below I assume $k = 1$, but in the more general case, k might be any positive number: the larger it is, the less the agent cares about the future.) Suppose, moreover, that the agent at $t = 0$ faces the choice between a reward of 10 at $t = 5$ and a reward of 30 at $t = 10$. At $t = 0$ the present value of the former is 1.67 and that of the latter is 2.73. An agent that maximizes present value will form the intention of choosing the delayed reward. At $t = 1$ the present value of the earlier reward is 2 and that of the later is 3. At $t = 2$ the values are 2.5 and 3.3; at $t = 3$ they are 3.3 and 3.75; and at $t = 4$ they are 5 and 4.29. At some time between $t = 3$ and $t = 4$, that is, the earlier reward ceases to be the least and becomes the most preferred option *as the result of nothing but the sheer passage of time*. It is easy to see, in fact, that the switch occurs at $t = 3.5$, which is when I call my dentist to cancel the appointment.

This pattern is even easier to see in a diagram. In Figure 6.3, the agent can either choose the small reward B at $t_1$ or wait until $t_2$ and get the larger reward A. The hyperbolic curves I and II represent the present values of these rewards as evaluated at various earlier times. They are in fact *indifference curves* (Chapter 10) that represent the trade-off between the time a reward becomes

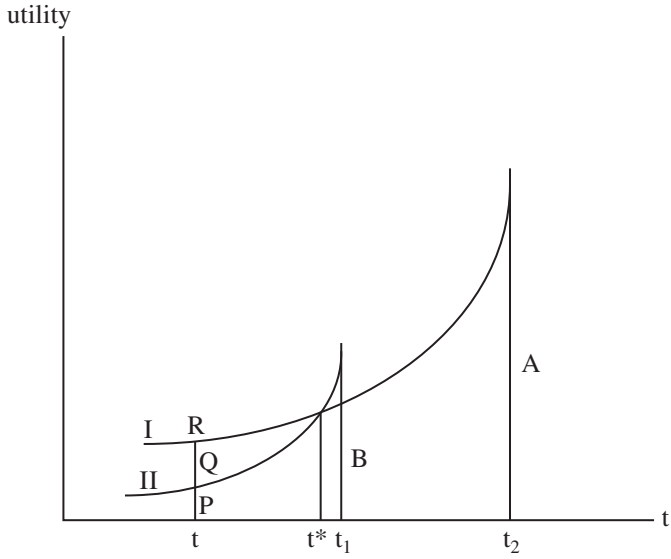



Figure 6.3

available and the size of the reward. At time t, for instance, the agent is indifferent between getting reward PQ immediately and getting the small reward at $t_1$, and also indifferent between getting PR immediately and getting the large reward at $t_2$. Since at time t the present value of A is larger than that of B, she will form the intention to choose A. Yet because the hyperbolic curves cross one another at t*, a preference reversal occurs at that time and she chooses B instead.[5]

[5] There is an alternative, slightly different way of representing hyperbolic discounting. It rests on the intuitive idea that people make a radical distinction between the present and all other times, by attaching more importance to welfare in the current period than to welfare in all later periods. In addition, they differentiate *among* later periods. In a three-period example, writing $u_i$ for experienced welfare in period i, the present value or discounted sum of utility is $u_1 + b(du_2 + d_2u_3)$. There are two discount factors involved. Compared to the present, all future utility, regardless of when it is experienced, is discounted by a factor b. In addition, all future utilities are discounted exponentially by a factor d. The present moment has a visceral salience that makes it stand out compared to all others, whereas later periods gradually lose their motivating power by something more akin to an optical illusion. This pattern, called "quasi-hyperbolic discounting," has in common with hyperbolic discounting proper that it can induce preference reversals. It differs in that the present value of an infinite stream of equal rewards (as in Pascal's wager) has a finite sum. There is some evidence from neurophysiology that quasi-hyperbolic discounting, although introduced only as a useful approximation to hyperbolic discounting, is in fact the more accurate representation.

### Pascal's wager

We can use Pascal's wager to illustrate the relation between exponential and hyperbolic time discounting. Pascal wanted to persuade the freethinking gamblers among his friends that they should bet on the existence of God, since even the smallest chance of eternal bliss would offset the greatest possible earthly pleasures. Pascal's argument harbors many complexities, some of which will concern us in the next chapter. Here I only want to draw attention to a question that Pascal does not mention: does the present (discounted) value of eternal bliss have a finite or an infinite value? If it is finite, the gambler might prefer to take his pleasures on earth rather than wait for the afterlife.

Suppose for simplicity that each period in the afterlife provides 1 unit of experienced utility; that the person expects to die in n years from the present; and, finally, that he discounts future welfare exponentially by a factor of k ($0 < k < 1$). If God exists and grants him salvation on the basis of his faith, the present value of bliss in the first year after he dies is $k^n$ units of utility, that of the second year $k^{n+1}$, and so on. As a matter of elementary algebra, this infinite sum ($k^n + k^{n+1} + k^{n+2} \ldots$) adds up to a finite sum $k^n/(1+k)$. Conceivably, at least, this sum might be inferior to the present value of n years of hedonistic living on Earth. By contrast, if the agent is subject to hyperbolic discounting the infinite sum $1/(n + 1) + 1/(n + 2) + 1/(n + 3) \ldots$ increases beyond any given finite value, implying that if we compare present values any earthly pleasure will ultimately be overtaken by the bliss of salvation. Even if the latter is multiplied by a small probability (as small as you wish) that God exists, the product will still increase beyond any finite number.

Suppose, however, that Pascal's interlocutor is regularly exposed to opportunities to gamble. When considered ahead of time, he prefers to attend mass rather than gamble, because the former will ultimately make him believe and assure him an expectation of infinite bliss. By the logic of hyperbolic discounting, however, the imminence of the opportunity to gamble will induce a preference reversal. He will form the intention to gamble just one more time and then start going to mass. With St. Augustine, he will say, "Give me chastity and continence, but not yet." Next week, the same reasoning will apply. Thus the very structure of time discounting that ensures that eternal bliss has the greater present value will also prevent the gambler from taking the steps to achieve it.

### Weakness of will

As this example shows, hyperbolic discounting may illuminate the problem of weakness of will. With the problem of self-deception (see next chapter), it

constitutes a classical instance of paradoxical irrationality. Both weakness of will and self-deception are forms of *motivated* irrationality, but that feature does not in itself make them paradoxical. Wishful thinking, too, is irrational, but hardly paradoxical. What lends an air of paradox to weakness of will and self-deception is that the person subject to them appears to *want and not want, believe and not believe, the same thing at the same time*. The paradox has led some thinkers and scholars to deny that these states can exist. Others have argued for their existence, and tried to specify mechanisms that can bring them about. My own view is agnostic, and somewhat skeptical.

A weak-willed (or *akratic*) person is characterized as follows:

1. The person has a reason for doing X.
2. The person has a reason for doing Y.
3. In the person's own judgment, the reason for doing X is weightier than the reason for doing Y.
4. The person does Y.

Emotions, in particular, are often held to have the capacity for inducing action against the better judgment of the agent. When Medea in Euripides' play is about to kill her children, she says, "I know indeed what evil I intend to do. But stronger than all my after thoughts is my fury." In Ovid's version of the play, she says, "An unknown compulsion bears me, all reluctant down. Urged this way or that ... I see the better and approve it, but I follow the worse."

These utterances, like the four statements used to characterize weakness of will, are all ambiguous or underspecified in that there is no mention of *when* they are supposed to be true. Let us define a *strict conception of weakness of will* as follows:

1. The person has a reason for doing X.
2. The person has a reason for doing Y.
3. The person does Y, judging *at the moment of action* that the reason for doing X is weightier than the reason for doing Y.

Imagine a person who has resolved to quit smoking and goes to a party where she is offered a cigarette. She accepts the offer, knowing *as she does so* that she should not. A person on a diet may accept an offering of dessert knowing *as he does so* that it is not a good idea. Although there is nothing impossible about this conception of weakness of will, it runs into two empirical problems. It would be hard to establish that the action and the "better judgment" coexisted at the very same moment, rather than that the judgment changed a split second before the action. Also, nobody to my knowledge has specified the causal mechanism by which the desire to do Y acquires greater causal efficacy than the desire to do X.

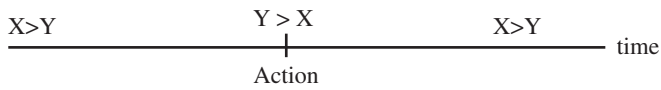



Figure 6.4

To bypass these problems we may define a *broad conception of weakness of will*, which allows the agent's judgment that he should do X and the choice of Y to occur at different moments:

1. The person has a reason for doing X.
2. The person has a reason for doing Y.
3. In the person's own calm and reflective judgment, the reason for doing X is weightier than the reason for doing Y.
4. The person does Y.

Socrates denied that weakness of will in the strict sense was possible. Aristotle, too, came close to suggesting the same thing. He allowed for WW in the broad sense, citing as an example a person whose judgment at the time of action is under the influence of alcohol. Suppose I go to the office party, have too many drinks, offend my boss, and make amorous advances to his wife. At the time, these actions seem the perfectly natural thing to do. Yet ahead of time, had anyone suggested I might act in this way, I would have rejected it as inconsistent with my calm, reflective judgment. If I had been persuaded that my judgment might dissolve in alcohol, I would have stayed away. After the fact, I might bitterly regret my behavior.

This case, shown in Figure 6.4, is a case of *temporary preference reversal*, not of weakness of will in the strict sense. There are at least three mechanisms that may bring about such changes. One is *temporal proximity*, as explained in the discussion of hyperbolic discounting. Another is *spatial proximity*, as illustrated by the phenomenon of cue dependence. This mechanism explains, for instance, many cases of relapse among addicts. Even after years of abstinence, an environmental cue traditionally associated with drug use may trigger relapse. Merely seeing drug paraphernalia on TV may be sufficient. The resolve to go on a diet may be undermined by the sight of the dessert trolley coming around. In these cases, too, the agent chooses according to her conception *at the moment of choice* of what she most prefers, all things considered. Finally, *passions* are capable of inducing temporary preference change, by virtue of the fact that they usually have a short half-life (Chapter 8). They may also induce preference reversal by causing the agent to pay less attention to the remote future.[6]

[6] In fact, the preference reversal caused by hyperbolic time preferences may be mimicked by emotionally induced changes in the discounting factor associated with exponential time

We may extend this idea to include temporary (and motivated) changes in the agent's *beliefs*. On this very broad conception, weakness of will can also result from self-deception (or wishful thinking). Having decided ahead of a party to have only two drinks in order to be able to drive home safely, a person might, under the influence of his desire for a third drink, tell himself, against the weight of the evidence, that it will not make a difference to his driving skills.[7] His preference for safe driving remains unchanged, but his belief about the conditions under which he can drive safely has changed. He might also, of course, undergo a temporary preference change, if he decides that having a good time at the party is so important that it offsets the risks (which he may perceive accurately) of drunk driving.

### Discounting the past

The impact of *expected* future events on present choices may depend, as we have seen, on the time at which the agent believes they *will* occur. The impact of *remembered* past events on present choice can depend on the time at which they *did* occur. Some past events, to be sure, fade from memory, but what is the impact of those that do not? Intuitively, we tend to think that recent events have a greater impact, partly but not only because they are more easily remembered. The intuition is confirmed by evidence that in presidential elections, American voters take account of changes in the nation's economic situation under the incumbent government mainly if they occur relatively close to election day, and are little affected by what happened in the first years of the administration.[8] They may be said to suffer from "backward-looking myopia."

Other facts suggest that the impact of the past is more ambiguous. When speaking before a jury, the prosecutor and the defense attorney have to decide whether to place their stronger arguments at the beginning of the presentation or at the end. Although the intuition just cited suggests that they should exploit "the recency effect" and place them at the end, there is also evidence for a

preferences. Suppose the agent faces the choice between two options, A and B, which offer the respective rewards (2, 5, 6) and (5, 4, 1) in three successive periods. With a one-period discounting rate of 0.8 (and a two-period rate of 0.64), the present values of the two options (as assessed in the first period) are respectively, 9.84 and 8.84. With a one-period discounting rate of 0.6 (and a two-period rate of 0.36), the values are 7.16 and 7.96. Not surprisingly, the agent ceases to prefer the option with the better long-term consequences when emotions cause him to pay less attention to the future.

[7] By contrast, if he is concerned with being stopped by the police rather than with having an accident, it is harder to make himself believe that the third drink will not cause the blood alcohol content to go beyond the legal limit. As I argue in the next chapter, even wishful thinking is (somewhat) subject to reality constraints.

[8] One would expect a rational and self-interested executive to exploit this fact, by boosting economic growth in election years. The evidence for the existence of a "political business cycle" remains ambiguous, however.

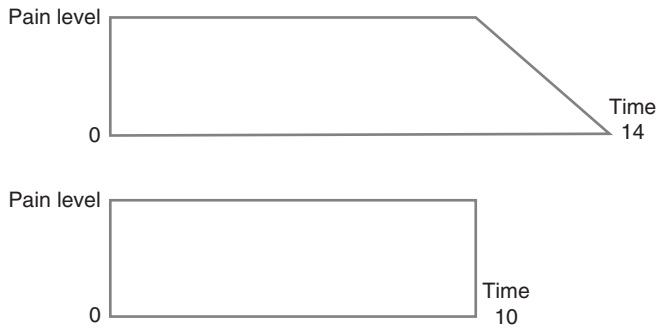



Figure 6.5

"primacy effect" that would support the opposite strategy.[9] Since both effects have been demonstrated, what should the speaker do? The so-called "Nestorian strategy" is to place the second-best argument at the beginning and the best at the end, with the least persuasive arguments in the middle.

Experiments suggest that we sometimes assess the past using a *"peak-end" heuristic*. In one experiment, each subject was exposed to a continuous loud and unpleasant sound in two conditions. In the condition shown in the top diagram of Figure 6.5, the sound was consistently loud for ten seconds and then tapered off for four seconds. In the second, the sound was cut off abruptly after ten seconds. When they were asked to choose which sequence they would prefer to experience again, about two-thirds chose the first – objectively more unpleasant – experience. A study where patients were assigned to two colonoscopy procedures that differed in the same way gave a similar result. The proposed explanation for these counterhedonic choices is that the subjects judged their experience (i) by how good or bad it was at its best or worst and (ii) how good or bad it was when it ended. Other experiments also showed "duration neglect": the duration of the experiences had little effect on how the subjects retrospectively assessed their pleasantness.

The peak-end effect and duration neglect have also been observed in the recollection of musical emotion, with an additional *slope effect*. According to the authors of this study, "The duration of . . . episodes contributes minimally to remembered affect (duration neglect). Listeners rely on the peak of affective intensity during a selection, the last moment, and moments that are more emotionally intense than immediately previous moments to determine post-performance ratings." This may well be true in general, but some pieces of

[9] *Pride and Prejudice*, whose draft title was *First Impressions*, testifies to the force of the primacy effect.

music, such as Beethoven's Fifth Symphony, certainly illustrate the primacy effect. Another question concerns the lessons that a composer might draw from these findings, assuming that they are robust. Should he maximize the experienced affect of the listeners or their remembered affect? If he chooses the latter, members of the audience may be more likely to listen to the piece again, although their experience may be inferior. If he chooses the former, may we charge him with paternalism?

## Motivational versus cognitive myopia

In the last section I made the obvious observation that agents cannot be influenced by their memory of the past if they do not have any memory of it. A similar point applies to the future. For future consequences of present choices to affect these choices, the agent must not only be motivated to take account of them, but also have the cognitive capacity to determine what those consequences will be.[10] Thus behavior that appears to be shaped by "motivational myopia" may, in reality, be shaped by a "cognitive myopia." Commenting on the suppression of the English monasteries under Henry VIII, Hume observed that "there is no abuse so great in civil society, as not to be attended with a variety of beneficial consequences; and in the beginnings of reformation, the loss of these advantages is always felt very sensibly, while the benefit, resulting from the change, is the slow effect of time, and is seldom perceived by the bulk of a nation."

Tocqueville argued that people in democratic societies are naturally myopic, in the sense of being unable to delay gratification. At the same time, he argued for the existence of a cognitive deficit: "what democracy often lacks is a clear perception of the future, based on enlightenment and experience." This argument applies with particular force to the benefits to be derived from liberty and the dangers that equality might bring.

> Political liberty, if carried to excess, can endanger the tranquility, property, and lives of private individuals, and no one is so blind or frivolous as to be unaware of this. By contrast, *it is only the attentive and clear-sighted who perceive the perils* with which equality threatens us, and they usually avoid pointing them out. They know that the miseries are remote and are pleased to think that they will afflict only future generations, *for which the present generation evinces little concern*. The ills that liberty sometimes brings on are immediate. They are visible to everyone, and to one degree or another everyone feels them. The ills that extreme equality can produce reveal themselves only

[10] Laboratory experiments rarely capture this fact, since subjects are *told* what the consequences of their choices will be, rather than having to determine them for themselves. Similarly, in experiments designed to capture risk attitudes, subjects are *told* the likelihood of the various outcomes. As a consequence, experiments may not allow us to distinguish between "motivational pessimism" (risk aversion) and "cognitive pessimism" (counterwishful thinking).

a little at a time . . . The goods that liberty yields reveal themselves only in the long run, and it is always easy to mistake their cause. The advantages of equality are felt immediately and can be seen daily to flow from their source.

The first phrase I italicize refers to a cognitive deficit in all citizens except for the "attentive and clear sighted." The second refers to a motivational deficit in (it would seem) all citizens. Even though the temporally distant effects of their present choices appear on the mental screen of members of the intellectual elite, they are not motivated to lend them any weight in their decisions. As for members of the majority, they lack foresight as well as prudence: the two deficits converge.

I argue in Chapter 8 that cognitive myopia can be induced by the urgency of emotion. It can also, however, simply be due to the increased thickness of the fog of uncertainty when we try to peer into the future. In such circumstances, decision makers are often tempted to base their choices on the short-term consequences they *can* foresee. The 2011 military intervention of Libya was motivated by the clear short-term benefits of getting rid of a dictator, with little thought about who or what would replace him. Similarly, an historian of the Vietnam War writes that "Fixated on short-term expedients and lacking a comprehensive estimate of what the war might cost the United States in the long term, [President Lyndon B. Johnson] focused on the more easily discernible price of withdrawal."

Many economic models assume that people have "rational expectations," in the strong sense that agents actually use the models to form their expectations. More weakly, the models assume that agents are not systematically wrong in forming their expectations. The strong version imputes to agents cognitive capacities that they demonstrably do not have. The weak version ignores the fact that some agents may not form any expectations at all, but recognize the pervasive uncertainty about what will happen a few years hence. It also ignores the fact that some agents may form *adaptive* expectations, as in the cobweb model further discussed in Chapter 17. These expectations can indeed be systematically wrong, but the agents may not be in a situation to learn from their mistakes. People also make other systematic predictive errors that are not corrected by experience. According to one analysis, the best strategy for goal keepers in face of a penalty kick is to stay in the center of the goal, yet in about 94 percent of the cases they throw themselves to one side or the other, perhaps because of "inaction-aversion" (Chapter 8).

### Bibliographical note

Rational-choice explanations of time discounting and altruism are offered by G. Becker and C. Mulligan, "The endogenous determination of time preference," *Quarterly Journal of Economics* 112 (1997), 729–58, and in

C. Mulligan, *Parental Priorities and Economic Inequality* (University of Chicago Press, 1997). For evidence that primates may be able to plan for future (not currently experienced) needs, see N. Mulcahy and J. Call, "Apes save tools for future use," *Science* 312 (2006), 1038–40. Two source books on time discounting and other aspects of intertemporal choice are G. Loewenstein and J. Elster (eds.), *Choice over Time* (New York: Russell Sage Foundation, 1992), and G. Loewenstein, D. Read, and R. Baumeister (eds.), *Time and Decision* (New York: Russell Sage Foundation, 2003). I discuss Pascal's wager at greater length in "Pascal and decision theory," in N. Hammond (ed.), *The Cambridge Companion to Pascal* (Cambridge University Press, 2004). The neurophysiological evidence for quasi-hyperbolic time discounting is in S. McClure *et al*., "Separate neural systems evaluate immediate and delayed monetary rewards," *Science* 306 (2004), 503–7. Modern discussions of weakness of will take off from D. Davidson, "How is weakness of the will possible?" in his *Essays on Action and Events* (Oxford University Press, 1980). I comment on his ideas in "Davidson on weakness of will and self-deception," in L. Hahn (ed.), *The Philosophy of Donald Davidson* (Chicago: Open Court, 1999). Motivated belief formation is discussed in D. Pears, *Motivated Irrationality* (Oxford University Press, 1984). I discuss the link between weakness of will and preference reversal at greater length in "Weakness of will and preference reversal," in J. Elster *et al.* (eds.), *Understanding Choice, Explaining Behavior: Essays in Honour of Ole-Jørgen Skog* (Oslo Academic Press, 2006). The evidence for "backward-looking myopia" in American politics is in L. Bartels, *Unequal Democracy* (Princeton University Press, 2010). Doubts about the political business cycle are expressed in A. Drazen, "The political business cycle after 25 years," in B. Bernanke and K. Rogoff (eds.), *NBER Macroeconomics Annual* 15 (2000), 75–137. The cited studies of the peak-end heuristic are D. Kahneman. P. Wakker, and R. Sarin, "Back to Bentham? Memories of experienced utility," *Quarterly Journal of Economics* 112 (1997), 375–406, and D. Redelmeier, J. Kart, and D, Kahneman, "Memories of colonoscopy: a randomized trial," *Pain* 104 (2003), 187–94. The study of the memory of music is A. Rozin. P. Rozin, and E. Goldberg, "The feeling of music past: how listeners remember musical affect," *Music Perception* 22 (2004), 15–39. The comment on Lyndon B. Johnson is from H. McMaster, *Dereliction of Duty* (New York: Harper, 1997), p. 297. The analysis of the actual and optimal behavior of goal keepers is in M. Bar-Eli *et al*., "Action bias among elite goalkeepers: the case of penalty kicks," *Journal of Economic Psychology*, 28 (2007), 606–21. Other studies yield different conclusions.

# 7 Beliefs

In the present chapter I discuss the causal history of beliefs, postponing to Chapter 13 a discussion of the normative principles of belief formation.

### What is it to "believe" something?

To understand the role of beliefs in generating action, we have to understand their nature, their causes, and their consequences. As I mentioned in the introductory remarks to Part II, it is not always clear what it means to "believe" that something is the case. Did the followers of Communism who "believed" that the party could do no wrong really *believe* it?[1] How can we tell the difference between the congenital pessimist who tends to believe the worst and the prudent decision maker who merely acts *as if* the worst-case scenario were true? How can we tell the difference between risk aversion (a formal preference) and pessimism (a belief)?

Also, in everyday language "belief" suggests less than full endorsement. I *believe* it will rain tomorrow, but I also know I might be wrong. I do not merely believe that I am married; I *know* it. In philosophical analyses, knowledge is usually defined as justified true belief, a belief that stands in a particular relation both to the world (it is true) and to the body of evidence the agent possesses (it is justified). Yet neither of these features of knowledge captures the subjective certainty that often underlies the phrase "I know" in ordinary discourse. This certainty is not simply the limit of 97 percent probability, 98 percent, 99 percent, 99.9 percent, and so forth. It is qualitatively different from anything short of certainty.

This "certainty effect" shows up in the following experiment. One group of subjects was asked to express their preferences over various options. (Numbers in parentheses indicate the proportion of subjects preferring a given option.)

[1] After the fall of Communism, a woman from the former East Germany said at a public meeting that her generation had been raised from childhood on to conform, to stay in line. A long-term schizophrenia had *hollowed them out* as people. So, this woman said, now she could not just suddenly "speak openly" or "say what she thought." She did not even really know precisely what she thought.

A 50 percent chance to win a three-week tour of England, France, and Italy (22 percent).
A one-week tour of England, with certainty (78 percent).

Another group was given the following options:

A 5 percent chance to win a three-week tour of England, France, and Italy (67 percent).
A 10 percent chance to win a one-week tour of England (33 percent).

Members of the first group tend to prefer the "England only" option because it is available *for sure*. Once it is deflated by the same probability as the alternative, the latter looks more attractive. Soldiers who are asked whether they will volunteer for highly dangerous missions may have disproportionately fewer hesitations than those who are asked to volunteer for suicide missions. The former may also, of course, be subject to wishful thinking ("It won't happen to me"), which has no purchase on the latter.

### Four cognitive attitudes

Even setting aside these problems, the idea of belief remains ambiguous. We may distinguish among four cognitive attitudes to the world, with decreasing strength. First is the mode of *certainty*. Second is the mode of *risk*, in which agents assign probabilities, whether based on past frequencies or their own judgment, to each of a set of mutually exclusive and jointly exhaustive outcomes. Third is the mode of *uncertainty*, in which people know the set of mutually exclusive and jointly exhaustive outcomes but find themselves unable to attach any (cardinal) probabilities to them. They may be able to attach *ordinal* probabilities to the outcomes, that is, to say that one outcome is more likely to occur than another, without being able to say *how* likely they are (see Chapter 13 for an example). For practical purposes, that situation is no better than uncertainty. Finally is the mode of *ignorance*, in which both the range of possible outcomes and their probability of occurrence are unknown or incompletely known.[2] In the memorable words of former Defense Secretary

[2] The state (or "veil") of ignorance may be *deliberately* induced, to prevent agents from acting according to their self-interest. If the electoral law is included in the constitution, one might specify that any changes will take effect only after n years + 1, where n is the length of the electoral cycle. Elected officials might be required to put their financial assets in a blind trust. A rare use of a veil-of-ignorance procedure in politics occurred when, in the words of an historian, Gaius Gracchus "compelled the Senate, which decided to which provinces consuls should be sent (normally after their year of office in Rome), to fix the provinces before the consuls were elected instead of during their consulships. Thus the Senate could less easily reward its favourites with the best provinces." He adds that "since the provinces would now have to be allocated eighteen months in advance, the new arrangement would not make for efficiency."

Donald Rumsfeld, we are facing not only known and unknown quantities, but also "unknown unknowns."

I focus on certainty and risk, not because these are always the appropriate cognitive attitudes, but because they are the most common ones. Even when people have no grounds for having *any* belief on a given topic, they often feel irresistibly compelled to form an opinion – not a specific opinion (as in wishful thinking), but *some opinion or other*. This propensity is to some extent determined by cultural factors. Albert Hirschman has said that most Latin American cultures "place considerable value on having *strong opinions* on virtually *everything* from the *outset*." In such societies, to admit ignorance is to admit defeat. But the tendency is really universal. Montaigne said that "many of this world's abuses are engendered – or to put it more rashly, all of this world's abuses are engendered – by our being schooled to be afraid to admit our ignorance and because we are required to accept anything which we cannot refute." The intolerance of uncertainty and ignorance flows not only from pridefulness, but from a universal human desire to find meanings and patterns everywhere (see Chapter 9). The mind abhors a vacuum.

Social agents are averse to uncertainty because it makes them feel uncomfortable. *Scholars* can be averse to uncertainty because it makes it difficult to prove theorems. The notion of expected utility maximization that is at the core of economic theory is undefined if expectations are. Economists, therefore, often attribute subjective probabilities to the agents without providing evidence that these have any kind of psychological reality. As I mentioned in the Introduction to Part II, they may for instance assume that in a situation of uncertainty, an agent will think all options equally likely (a "uniform" or flat probability distribution). As I argued, this procedure is arbitrary. Natural scientists also appeal to uniform distributions when trying to understand natural phenomena, for instance when trying to predict climate change. One study criticizes this procedure by observing that "physically, we can equally well use a parameter labeled 'ice fall rate in clouds' or its inverse ('ice residence time in clouds') and achieve identical simulations. Sampling uniform distributions under each of the two different labels however, yields completely different results."

In other cases, scholars assume that agents have rational expectations (Chapter 6), that is, that they share the information and risk assessments of the modeler. Often, this procedure is manifestly adopted because of the need to prove theorems rather than because it is supported by evidence. In the Conclusion, I cite a particularly egregious example in which scholars impute to revolutionary agents a sharp probability about the future state of the economy. After the recent financial crisis, however, many economists have been more willing to follow Keynes in his rejection of the idea that agents act to maximize "a weighted average of quantitative benefits multiplied by quantitative probabilities."

These error-generating mechanisms rely in one way or another on *motivation*. Yet error can also arise from *ignorance*. The point seems obvious but is actually a bit subtle. Darwin noted, for instance, that "ignorance more frequently begets confidence than does knowledge." Ignorance together with confidence is a good recipe for error. Conversely, when the circle of light expands, so does the surrounding area of darkness, inducing greater humility. As Adam Smith noted, it "is the inferior artist only, who is ever perfectly satisfied with his own performances." Experiments suggest in fact that incompetence not only causes poor cognitive performance, but also the inability to recognize that one's competence is poor. The incompetent are doubly handicapped.

### Subjective assessments of probability

Probability judgments can stem from observation of objective frequencies or be purely subjective evaluations.[3] When the agent can draw on a large number of observations of similar situations, the frequentist method can yield good results. If I plan to have a picnic on my birthday next month and need to form an opinion about the likely weather, the best I can do is probably to look up the weather statistics for the same day in previous years. But if I need to form an opinion about the weather tomorrow, the best single predictor is today's weather. It is not, however, the only predictor. Past records can tell me whether sunny weather on that day is a rare or normal event. If it is rare, today's sunny weather loses some of its predictive value. I may consult the barometer on my wall to see whether the air pressure is rising or falling or look at the evening sky, the flight of the swallows, and so on.

To integrate all this information into an overall probability judgment about tomorrow's weather is a difficult task. Most of us are not very good at it. Often, the problem is not lack of information, but an abundance of it, combined with the lack of a formal procedure for integrating it into an all-things-considered opinion. Some people, though, are better than most of us at integrating vast and diffuse information with varying degrees of relevance into an overall assessment. They possess the elusive but crucial quality of *judgment*. Successful generals, businesspeople, and politicians tend to have it – that is why they succeed. A good central banker needs to have it, but most economists do not.[4]

[3] In a deeper analysis, the first (objective) method boils down to the second (subjective) one, since to be useful objective data always need a subjective interpretation. For many practical purposes, though, the distinction is clear and useful.

[4] Writing about Alan Greenspan (*New York Times*, October 28, 2005), Paul Krugman noted that while distrusting formal models he had "the ability to divine from fragmentary and sometimes contradictory data which way the economic wind was blowing." In his more recent writings, subsequent to the financial crisis, Krugman has been more critical of Greenspan.

The best the rest of us can do is to recognize that we do not have it and learn not to trust our intuition. I may learn, for instance, that I often distrust people for reasons that, when I come to understand them, are irrelevant. ("He looked like a bully I knew in fifth grade.") Hence I may come to distrust my distrust.[5]

We tend to think, however, that judgment is possessed not only by successful generals, politicians, and businesspeople, but also by trained experts. In complicated matters of diagnosis or prognosis, such as identifying psychotic individuals or assessing how likely it is that a person who requests early release from prison will commit a second offense, we trust the expert. Because of their experience, experts are sensitive to telltale signs that untrained observers might ignore or whose significance they might not understand. Moreover, when different pieces of evidence point in different directions, experts can draw on their experience to decide which, in any given case, should be given most weight. This at least is how we think about experts. As most of us consider ourselves experts in some domain or other, if nothing else in predicting the behavior of our boss, spouse, or children, we have a great deal invested in this image of the superior cognitive skills of the expert.

Unfortunately, *this image is thoroughly false*. In many studies the diagnostic or prognostic performance of experts has been compared with the performance of a simple mechanical formula based on a few variables. Essentially, this amounts to comparing objective (frequentist) methods and subjective ones. The weights assigned to the variables are derived by statistical techniques that assign the weights most likely to predict observed outcomes. Almost without exception, the formula performs at least as well as the expert and usually better.[6] In a study of the diagnosis of progressive brain dysfunctioning based on intellectual testing, to cite only one example, a formula derived from one set of cases and then applied to a new sample correctly identified 83 percent of the new cases. Groups of experienced and inexperienced clinicians correctly identified 63 percent and 58 percent, respectively. Moreover, experts often disagree strongly with one another. In another study, highly experienced psychiatrists who viewed the same psychiatric interview could not agree on the patient's diagnosis, motivations, or feelings. Some psychotherapists use responses to ambiguous inkblots as cues to diagnoses. It appears, however, that the patients are as ambiguous to them as the inkblots to the patients.[7]

[5] Knowing that one may be subject to bias is one thing; being able to correct it is another. Studies show that deliberate attempts to debias one's judgment are of little value, since one easily falls into the traps of insufficient correction, unnecessary correction, or overcorrection. One may learn to distrust one's judgment, but it is harder to improve it. If one were able to, there might be no need to.

[6] This superiority remains even if we simply assign equal weights to all variables!

[7] In the 2012 trial of the Norwegian mass murderer Anders Breivik, the first team of psychiatrists concluded that he was a "paranoid schizophrenic," who could not be held legally responsible for

### Some errors of statistical inference

Experts no less than laypersons often go wrong because they ignore obvious or not-so-obvious principles of statistical reasoning. In one study, subjects were given a description of a young man with long hair and a habit of reading poetry and asked whether they thought it more likely that he was an orchestra violinist or a truck driver. Most said he was more likely to be a violinist, thus ignoring the *base rate* of the two groups, that is, the absolute number of individuals in each. There are so many more truck drivers than orchestra violinists in the nation (and so much variation among truck drivers) that the poetic young man is in fact more likely to drive a truck. This mistake is referred to as the *base-rate fallacy*.

Another source of mistakes in belief formation is *selection bias*. Think of a man arriving at a railway station and examining a map of the local area with a large red dot on it labeled "You Are Here," and being amazed that the railway company knew he would be there at that time. More seriously, patients in dialysis centers are often surprisingly reluctant to be on the waiting list for a kidney transplantation. One reason is that all the transplanted patients they ever see are those for whom the operation failed so that they had to go back on dialysis. Montaigne cited a bias of this kind when he referred to Diagoras as being "shown many vows and votive portraits from those who have survived shipwrecks and … then asked, 'You, there, who think that the gods are indifferent to human affairs, what have you to say about so many men saved by their grace?' – 'It is like this,' he replied, 'there are no portraits here of those who stayed and drowned – and they are more numerous!'" Similarly, a psychiatrist who claims that "no child abusers ever stop on their own" neglects the fact that if any do he is unlikely to have met them. In *À la recherche du temps perdu* the Narrator observes about Charlus that "everything made him become Germanophile, because … he lived in France. He was very keen-witted and in all countries fools outnumber the rest; no doubt, if he had lived in Germany the German fools defending an unjust cause with passion and folly would have irritated him; but living in France, the French fools, defending a just cause with passion and folly, irritated him no less." As a final example of this very common reasoning flaw, consider an inference by American officers during World War II: observing that there were fewer bullet holes in the engines of planes that came back than in other parts, they concluded that armor needed to be concentrated in what appeared to be the most vulnerable parts. The statistician Abraham Wald observed, however, that the reason

his actions. Following a public outcry, the court appointed a second team, which concluded that Breivik was a right-wing fanatic who could be held responsible. In her summing-up, the judge agreed with the second team. Public confidence in forensic psychiatry was shattered.

planes were coming back with fewer hits to the engine was that most planes that got hit in the engine were not coming back. In the words of the writer from whom I take this example, "The armor goes where the bullet holes aren't" rather than where they are.

Israeli air force leaders made a less obvious mistake when assessing the relative efficacy of reward and punishment in the training of pilots. Noting that the performance of pilots improved when they were punished for a bad performance but not when they were rewarded for a good one, they concluded that punishment was more efficient. In doing so, they ignored the phenomenon known as *regression to the mean*. In any series of events that are fully or partly determined by chance, there is a tendency for an extreme value on one occasion to be followed by a less extreme value on the next. Tall fathers get sons who are shorter than they are, and bad pilot performances are followed by less bad ones, independently of reward and punishment. When athletes who have done exceptionally well in one season do less well the next, fans and coaches often say they have been spoiled by success, when what we observe may only be regression to the mean.

Many scares about incidence of disease or harm in a particular community or profession are due to what epidemiologists call the "Texas sharpshooter effect": blast a barn door with a shotgun and then find the holes that are closest together. Draw a target around them and it looks like you hit a bull's-eye. When there was a cluster of suicides in the French postal service some years ago, newspapers blamed a toxic work environment, until statisticians pointed out that any random process will generate clusters (see below). Many cancer scares are also spurious. A specialist on epidemiology calculated that given a typical registry of eighty different cancers, you would expect 2,750 out of California's 5,000 census tracts to have statistically significant but perfectly random elevations of cancer. In Chapter 9, I return to the causes and effects of such spurious pattern finding.

The gambler's fallacy and its (nameless) converse offer another example. The purchase of earthquake insurance increases sharply after an earthquake but then falls steadily as memory fades. As do gamblers who make the mistake of believing that red is more likely to come up again if it has come up several times in a row, the purchasers form their beliefs by using the *availability heuristic*. Their judgment about the likelihood of an event is shaped by the ease with which it can be brought to mind, and recent events are more readily available than earlier ones. The decay of emotion over time (Chapter 8) might also be a factor. Conversely, people living in areas that are subject to frequent floods often believe that a flood is less likely to occur in year n + 1 if one has occurred in year n. As do gamblers who make the mistake of believing that red is less likely to come up again if it has come up several times in a row, they form their beliefs by relying on the *representativeness heuristic*. They believe,

or act as if they believe, that a short sequence of events is likely to be representative of a longer sequence in which it is embedded.

People often fail to grasp the relation between random processes and the distribution of outcomes. During World War II, many Londoners were certain that the Germans systematically concentrated their bombing in certain parts of their city, because the bombs fell in clusters. They did not understand the basic statistical principle that random processes tend to generate clustering, and that bombs falling in a neat gridlock pattern would have been stronger evidence of deliberate target selection. A fact that never fails to surprise those who have not come across it before is that in a group of as few as twenty-three people, the probability that two of them have the same birthday (day and month) is more than 50 percent. Out of a thousand investment firms, thirty are statistically likely to offer good advice on five successive occasions even if they basically perform the equivalent of tossing a coin. Clients who invest with these firms on the basis of their past performance, wrongly understood as the result of skill rather than luck, may experience that simple induction can lead one astray.

### Magical thinking

Consider next various forms of *magical thinking*, that is, the tendency to believe one can have a causal influence on outcomes that are actually outside one's control. Many readers, afraid of tempting fate, will have had the thought, "If I don't take my umbrella, it's sure to rain." People will also place larger bets on a coin that has not yet been tossed than on a coin that has already been tossed and for which the outcome has been concealed. In Proust, the Narrator's friend Robert Saint-Loup was subject to "a sort of superstitious belief: that the fidelity of his mistress to him might depend on his to her." (It is clear from the context that he was not referring to a causal influence.) Also, people may fail to grasp the distinction between *causal and diagnostic relevance*. In one experiment, subjects who were led to believe that the length of time they could hold their arms in painfully cold water was the best indicator of longevity held their arms in the water longer than those not given this (false) information.[8] Also, using their own behavior as a predictor of how others will act, people may

[8] This distinction between cause and symptom is not always evident. As late as 1959, the great statistician R. A. Fisher, assuming a genetic trait that predisposed the individual both to smoking and to cancer, argued that smoking was diagnostic of lung cancer rather than its cause. (It is true that he was in the pay of tobacco companies at the time.) Or consider the finding, discussed in Chapter 2, that the longer an individual has been out of work the less likely it is that he will find a job in a given time span. The duration of unemployment might be simply diagnostic of employability, or it could make a causal contribution (through demoralization, etc.) to the chances of finding employment.

choose the cooperative strategy in a Prisoner's Dilemma as if they could somehow bring it about that others cooperate too. In one experiment, cooperating subjects who were asked to predict the choice of their interaction partner as well as that of a non-partner who was matched with another person were more likely to predict (and had greater confidence in their prediction) cooperation by their interaction partner than the non-partner.[9] Public authorities sometimes seem to count on the susceptibility of citizens to magical thinking. Thus in Paris buses one finds a sign saying: "Qui salit le siège à l'aller risque de se tâcher au retour" (if you dirty the seat going out, you risk getting stained coming back).

Calvinism offers an example of this kind of magical thinking (Chapter 3). Given the Calvinist belief in predestination, there would seem to be no reason for a Calvinist not to indulge in all sorts of worldly pleasures, which by assumption cannot affect their fate after death. Max Weber claimed that Calvinism nevertheless made its followers adopt an ascetic lifestyle, not to gain salvation but to acquire the subjective certainty of being among the elect. We may read him as saying that the Calvinists confused the causal and diagnostic relevance of their behavior. This is made quite explicit in a letter circulated by English Baptists in 1770: "Every soul that comes to Christ to be saved . . . is to be encouraged . . . The coming soul need not fear that he is not elected, for none but such would be willing to come." We may think of this as retroactively forcing the hand of God, as opposed to prospective forcing by good works.

These errors (and many others that have been extensively documented) are for the most part "cold" or unmotivated mistakes, similar in some respects to optical illusions. Other errors, or "hot" mistakes, arise because the beliefs of the agents are *motivated*, that is, unduly influenced by their desires. As we shall see in Chapter 13, a causal influence of desires on beliefs is not intrinsically irrational. A desire can provide a reason for investing a specific amount of resources in information acquisition. The information thus obtained may serve as a reason for holding a certain belief. Although the desire does not provide a reason for holding the belief, it enters into a rational complex of belief formation. What drives in the wedge between the initial desire and the final belief is the fact that the outcome of the search for information is, by definition, not known at the time the decision to search is made.

### Motivated belief formation

The direct influence of desires on beliefs I just cited is, uncontroversially, consistent with rationality. A more controversial idea was provided by Pascal's

[9] This discrepancy allows us to exclude that the imputation of cooperative behavior to the interaction partner could have been due merely to the "false consensus effect" (Chapter 22).

wager. As I explained in the last chapter, Pascal argued that an agent who believes that there is a non-zero probability, however small, that God exists, should for the purely instrumental reason of maximizing expected value try to acquire a firm belief (in the mode of certainty) that God exists because, if he does exist, that belief will ensure eternal bliss. The premises for the argument are (1) that certain belief is certain to provide salvation and (2) that the instrumental origin of the belief does not detract from its efficacy for salvation. Although both premises may be dubious from a theological point of view, especially as Pascal also believed in predestination, this need not concern us here. The question is whether this "decision to believe" is a rational project. In one sense it is not: I cannot decide to believe at will the way I can decide to raise my arm at will. One might, however, use an indirect strategy. By acting *as if* one believed, Pascal argued, one will end up believing. The mechanism by which this might happen is, however, somewhat unfathomable.

There are other cases in which one might want to acquire a belief one believes to be false, because of the good consequences of holding it. If I want to cut down on my drinking but find myself insufficiently motivated by the risk of becoming an alcoholic, I may desire to believe that the risk is larger than I now believe it to be. By and large, however, there is no reliable technology for acquiring such beliefs. Unless the process has a *self-erasing component*, by which the origin of the belief in the desire to acquire it is eliminated from the conscious mind, the desire is likely to remain a mere wish.

In the "uncontroversial" case, the agent's desire induces a certain level of information gathering that will in turn induce some belief or other. In the "controversial case" the desire induces specific behavior that will in turn induce a specific belief the agent wants to hold. Both are indirect strategies. I now turn to beliefs that are *directly* shaped by motivation. This can come about in one of two ways, corresponding to two basic features of motivations: arousal and content. Just as we say that the stone broke the ice by virtue of its weight, not of its color, we may say that a motivation affects belief not by virtue of its content, but by virtue of the accompanying arousal level. Moderate physiological arousal can improve the quality of belief formation, by focusing attention and stimulating the imagination. "When a man knows he is to be hanged in a fortnight," Dr. Johnson said, "it concentrates his mind wonderfully." Beyond a certain level of arousal, however, cognition deteriorates. In states of extreme hunger, stress, fear, or addictive craving, it is hard to think straight because the arousal makes it difficult to keep previous reasoning steps in mind. Presumably, mental concentration is blunted when the hanging is but one day away. In scholastic aptitude tests a very strong motivation to get it right may actually cause one to get it wrong, just as a shooter's strong desire to hit the target may cause her hands to shake so that she misses (see Chapter 10). In the next chapter I argue that because of the urgency of many emotions, they

may cause the agent to bypass the normal machinery of rational belief formation. Thus beliefs may be *shaped by motivation* yet not be *motivated*, because the agent has no particular desire to believe they are true. Arousal *clouds* the mind but does not *bias* it in favor of any particular belief.

Motivated beliefs are of two main varieties. As I noted earlier, the agent may be motivated to hold *some belief or other* on a given topic, because of a need for closure or an intolerance of admitting ignorance. Alternatively, he may be motivated to hold some *specific* belief, such as the belief that his spouse is being faithful to him. The most important mechanisms generating this variety are rationalization, wishful thinking, and self-deception.

### *Rationalization*

Rationalization can provide the agent with a belief that justifies her mistakes or serves as a reason for doing what she would want to do anyway. The second mechanism is similar to transmutation (Chapter 9), and might also have been discussed under that heading.

*Rewriting the past* in order to absolve oneself (or one's affiliates) from blame is extremely common. In divorce proceedings, the two spouses often try to shift the blame for the breakdown of the marriage on the other. They may produce motivated and incompatible narratives, for instance by claiming percentages of time spent on housework or with children whose sum exceeds 100 percent. At the other end of the spectrum of importance, citizens of Germany, France, England, Austria, Serbia, and Russia tend to blame another country than their own for the outbreak of World War I.[10] A common mechanism is what one might call "reverse hindsight bias." When a risky choice fails, *observers* often claim that the failure was foreseeable (the choice had negative expected value). Even when the claim is inaccurate, it may be supported by hindsight bias. The *decision maker* might claim that it "looked like a good idea at the time" (the choice had positive expected value). Even when that claim is inaccurate, it may be supported by rationalization. Thus when it became clear that the critics of the Vietnam War had been right from the beginning, defenders such as Walt Rostow rationalized it by redefining the goal as "buying time" or creating "breathing room" for other countries in the region.

A very common form of rationalization is framing self-interested behavior as disinterested. In Jane Austen's *Persuasion*, we find an exchange between Sir Walter Elliot, one of the most finely drawn egotists in fiction, and an

[10] The idea that *no* single state might be to blame is undermined by the need to find meaning and order in the universe, specifically by the agency bias (Chapter 9). Even if this bias is overcome, the tendency to absolve one's own country might persist.

unnamed female interlocutor. He asks how his daughter Mary is doing, adding that "the last time I saw her she had a red nose, but I hope that may not happen everyday." Upon being reassured that his daughter was in very good health and very good looks, he responds that "If I thought it would not tempt her to go out in the sharp winds, and grow coarse, I would send her a new hat and pelisse." The heroine of the novel, his daughter Anne, considered "whether she should venture to suggest that a gown, or a cap, would not be liable to any such misuse." To rationalize his reluctance to spend money on his daughter, Sir Walter limits his options to gifts that he can reject as not being in her objective interest. Similarly, stingy parents may rationalize their refusal to help their children out financially by claiming that it will detract from their incentive to work, and rich nations may use the same argument to justify their refusal to aid poor countries. The fact that these arguments are in the interest of those who make them does not, of course, prove that they are wrong or are motivated by self-interest. To establish such claims, more evidence would be needed.

### *Wishful thinking*

Let me turn to wishful thinking and self-deception. These two ill understood phenomena have in common that a desire that *p* be the case causes the belief that *p* is the case. In wishful thinking this is a simple one-step process: the wish is the father of the thought. The evidence is not so much denied as ignored. As a result, the wishfully formed belief might happen to be the very same one that would be justified by the evidence, had it been consulted.[11] Self-deception as usually conceived involves four steps: first, the evidence is considered; second, the appropriate belief is formed; third, this belief is rejected or suppressed because it is inconsistent with our desire; and last, the desire causes another and more acceptable belief to be formed in its place. Self-deception is a

[11] Ignoring this point could be a source of irrational belief formation. Because it is often easy to detect the operation of motivated belief formation in others, we tend to disbelieve the conclusions reached in this way, without pausing to see whether the evidence might in fact justify them. Until around 1990 I believed, with most of my friends, that on a scale of evil from 0 to 10 (the worst), Communism scored around 7 or 8. Since the recent revelations I believe that 10 is the appropriate number. The reason for my misperception of the evidence was not an idealistic belief that Communism was a worthy ideal that had been betrayed by actual Communists. In that case, I would simply have been victim of wishful thinking or self-deception. Rather, I was misled by the hysterical character of those who claimed all along that Communism scored 10. My ignorance of their claims was not entirely irrational. On average, it makes sense to discount the claims of the manifestly hysterical. Yet even hysterics can be right, albeit for the wrong reasons. Because I sensed and still believe that many of these fierce anti-Communists would have said the same regardless of the evidence, I could not believe that what they said did in fact correspond to the evidence. I made the mistake of thinking of them as a clock that is always one hour late rather than as a broken clock that shows the right time twice a day. Later, I made the same mistake about members of the ecology movement.

paradoxical phenomenon, whose existence and even possibility have been called into doubt, so let me begin with the simpler issue of wishful thinking.

Before suggesting a mechanism by which wishful thinking is brought about, let me first state that, unlike what is the case for self-deception, it is impossible to deny its existence. One may deny that it occurs in high-stake situations or that it affects aggregate behavior such as stock markets or elections, but not that it occurs. If nothing else, world literature would testify to its existence. Moreover, many wishfully formed beliefs serve as premises for *action*, and hence are more than mere "quasi-beliefs." Some smokers who fool themselves into believing that smoking is not dangerous, in general or for them specifically, would have quit or tried to quit had they held more rational beliefs.[12] Overconfident individuals, who wishfully believe they are more capable than they really are, may embark on ventures they would otherwise have avoided. People who fool themselves into thinking they are as successful as others may lose a spur to improve themselves. A common mechanism is the following. First, a person is motivated to believe he is successful. Second, he finds some areas in his life in which he does in fact do well. Third, he enhances the importance of those areas to be able to tell himself that he is successful overall. Finally, he relaxes his efforts to succeed in other walks of life.

To navigate in life, it is instrumentally useful to have accurate beliefs. At the same time, beliefs may be intrinsically pleasant or unpleasant, that is, cause positive or negative emotions. If told that I have cancer, I can seek treatment, but the belief will also make me feel horrible. In Freud's language, those governed by the reality principle seek accurate beliefs, whereas those subject to the pleasure principle seek pleasant beliefs. This distinction applies only to beliefs in the strict sense, not to quasi-beliefs. People who form unrealistic beliefs about receiving a big monetary prize for their achievements yet do not spend the prize money before they have received it are at worst subject to a harmless form of the pleasure principle. In the more noxious variety, their conviction that they will receive the prize actually causes them to go into debt. The institutions of *tontines* and *rentes viagères*, in which one person receives a sum of money or a property upon the death of an unrelated person, probably owe some of their popularity to wishful thinking (and their notoriety to the crime novels in which they are part of the plot).

Wishful thinking cannot produce just any kind of pleasant belief, as it is somewhat subject to *constraints*. An agent who begins smoking may be tempted to form the wishful belief that smoking is not dangerous, or at least not dangerous for her. In doing so, however, she may be constrained by her prior beliefs about the dangers of smoking. The first time a person does badly

[12] As in the case of alcohol, quitting might require the irrational belief that smoking is *more* dangerous than it actually is.

on an exam, he may tell himself a story about bad luck, but if the same outcome occurs on the next four occasions the story is less likely to work if he fails for a sixth time. Or consider the example of the expensive Broadway show tickets that I introduced in Chapter 1. If I have paid $75 for the ticket but the show is lousy, my recollection of what I paid is likely to be too vivid to be subject to wishful downward revision. Given the intangible and multidimensional nature of aesthetic appreciation, it is easier to adjust my evaluation of the show upward. Similarly, although there is evidence both that likely events are seen as more desirable and that desirable events are perceived as more likely, the latter effect is more heavily constrained than the former.

In an instructive experiment, subjects expected to participate in a history trivia game with a given person either as their partner or as their opponent. After exposure to a sample of the person's performance, in which he got a perfect score, those who expected the person to be their partner (and therefore wished him to have high ability) judged him as better at history than those who expected him to be their opponent (and who therefore wished him to have low ability). At the same time, subjects were clearly constrained by the nature of the information they received, since even subjects expecting him to be their opponent judged him as better than average. A limitation of the experiment is that it did not offer the subjects an opportunity to *act* on these beliefs, with the potentially costly consequences that might follow from underestimating an opponent. For all we know, they might be mere quasi-beliefs. If they had been playing for money, the subjects might have been more circumspect.

In the examples just given, wishful thinking is constrained by prior factual beliefs. In other cases, it may be constrained by plausible causal beliefs. Wishful thinking often involves "telling oneself a story," the idea of a story being closely related to the idea of a mechanism that I discussed in Chapter 2. The plethora of mechanisms makes it easier to find some story or other that will justify any belief one might want to be true. I may dismiss an unwelcome rumor with the proverb "Rumors often lie" and embrace a welcome one by the proverb "Rumors rarely lie." Or suppose I read in the application material for a school of social work that emotional stability is highly desirable for people in that profession. If my mother left the workforce to take care of me when I was born, I may bolster my belief in my stability by telling myself a story that children benefit from the full-time attention of their parents. If she kept her job and sent me to day care, I may instead adopt a story that children benefit from being with other children and from having parents who have professional fulfillment outside the home.[13] If my favorite soccer team does badly, I can maintain my belief in its superiority if the other team won by (what can be

[13] As a matter of fact, no consistent differences are found in the later development of children brought up in these two environments.

construed as) a fluke event. "If the ball hadn't been deflected by the referee, the wing player would have received the pass in a position to score." If my horse finishes second, I can maintain my belief in my betting skills by saying that it "almost won." In an even more blatantly irrational piece of wishful thinking, if I put money on 32 and 33 comes up, I can also say that I "almost won" even if the two numbers are far from each other in the roulette wheel.[14]

Sometimes, however, there is no readily available and plausible story. Suppose a person places his money on 24. The number that did come out was 15, which is adjacent to 24 on the number wheel; hence his belief in his gambling skills is confirmed. Probably he would have considered other outcomes, such as 5, 10, and 33, also confirmations, because they are nearby on the wheel. Also he could have taken the outcomes 22, 23, 25, and 26 as confirmations because their numerical value is closer, or the numbers 20, 21, 26, and 27 because they are adjacent on the table. Thus 13 out of 37 possible outcomes could be taken as confirmations of his betting ability. But that also means that there are 24 outcomes for which no simple story is available. If one of these occurs, even a person prone to wishful thinking and highly motivated to adopt a specific belief might have to face the facts.

### *Self-deception*

Consider now the thorny issue of self-deception. The canonical definition is:

1. The individual holds two contradictory beliefs (that *p* and that *not-p).*
2. These two contradictory beliefs are held simultaneously.
3. The individual is not aware of holding one of the beliefs.
4. The act that determines which belief is and which belief is not subject to awareness is a motivated act.

The simultaneity condition is similar to condition (3) in the strict definition of weakness of will (Chapter 6). If these conditions are relaxed, both self-deception and weakness of will lose their paradoxical character. In Chapter 6, I argued that if we adopt a broader definition of weakness of will, actions that appear to go against the better judgment of the agent may simply be due to a preference reversal. In the case of self-deception, the paradox would disappear if the motivated adoption of the belief that *not-p* went together with the *erasure* from the mind of the belief that *p* rather than its relegation to the unconscious. Although this situation may be frequent, I shall focus on the full-blown paradoxical case. Standard examples include the denial that one's spouse is having an affair, that one is exhibiting the

[14] At the same time, people may be more disappointed if their number is close to the winning number. Some national lotteries offer small "consolation prizes" to those who "almost won."

symptoms of cancer, that one is gaining weight, that one is drinking more than x glasses of wine per day, and so on.

Before I proceed, let me comment on the ambiguous role of the desires (the motivation) in self-deception. Does the suspicious husband desire that his wife *be* faithful, or desire to *believe* that she is? He might in fact want her to be unfaithful, to justify his belief, but also want his belief to be false, to save his marriage. Although this example may be contrived, the following is not. A person who predicts that something he wants not to happen is nevertheless going to happen (a war, an earthquake) might be reluctant to give up his belief even when disconfirming evidence comes to light, because his amour-propre would suffer, but at the same time might welcome evidence that disaster will not strike. In other words, people make a *psychic investment* in their beliefs that, independently of their content, makes them reluctant to face the facts. As La Rochefoucauld said, "Nothing blights our self-esteem so much as to disapprove of what we once approved." Here I ignore this complication, to which I return in Chapter 9.

To my knowledge, experimental psychologists have not tried to verify the existence of weakness of will, in the strict sense, in the laboratory. By contrast, they have tried to demonstrate the existence of self-deception, in the strict synchronic sense. I shall discuss four such studies, and conclude by discussing some literary treatments of self-deception.

The authors who proposed the "canonical definition" of self-deception cited earlier also offer a putative example of the phenomenon. Their experiments turned on the belief of subjects as to whether a voice recording was of their own speech or of another person's. Subjects with a poor self-image tended to attribute their voice to another person, just as some people avoid looking at themselves in a mirror; at the same time, their galvanic skin response was stronger when they heard their own voice than when they heard that of another person. These facts suggest that conditions (1) and (3) of the definition are satisfied. Condition (2), simultaneity, was ensured by the experimental set-up. The satisfaction of condition (4), the motivated character of the non-awareness that they were hearing themselves, was assured by the finding that subjects who tended to ascribe the voices of others to themselves scored highly on a narcissism score. Although the experiments were more complex and subtle than I have indicated, my simplifications do not affect the objection that the experiment did not prove that the unconscious response *had a propositional basis*. One cannot exclude that the subjects reacted with *disgust* upon hearing their own voice, without formulating to themselves the proposition "This is my voice." Disgust is, in fact, one of the few emotions that do not require a propositional trigger (see Chapter 8).

The same comment applies to another experiment that was not designed to prove the reality of self-deception, but has been used as an argument for its

existence. As background, let me cite a remark put in the mouth of Roy Cohn, the legal adviser of Joseph McCarthy, in the play *Angels in America:* "Roy Cohn is not a homosexual. Roy Cohn is a heterosexual man, who fucks around with guys." This combination of homosexuality and homophobia suggests self-deception. In the experiment, homophobic and non-homophobic men were exposed to homosexual stimuli. The former were significantly more aroused, as measured by changes in the circumference of the penis. Their self-reports of arousal thus measured were low, however, and not different from those of non-homophobic subjects. Assuming that the self-reports were sincere, these findings might seem to provide evidence for self-deception: the conscious belief of the homophobic subjects that they were not turned on by homosexual material was inconsistent with their bodily responses. Yet once again, the findings do not provide evidence of an unconscious *propositional* belief about their homosexual tendencies, since the cause of their denial may have been disgust at their arousal.

I have already referred to the third experiment, in which some subjects were told that the time they could tolerate holding their hand in very cold water was an indicator of a heart condition predicting their life expectancy. They tended to hold their hands in the water longer than subjects who were not given this (false) information. These facts, together with self-reports, suggest the presence of self-deception. On the one hand, the subjects could hardly hold the conscious belief that they could modify the cause by acting on the symptoms. On the other hand, their behavior seemed to reveal that they unconsciously did hold that magical belief. Also, they were clearly *motivated* to keep the belief unconscious, since otherwise they could not, as most of them did, form and report the gratifying belief that their heart condition and life expectancy were indeed good. In this experiment, then, the proposed evidence for a repressed belief was behavior, not, as in the previously discussed experiments, a somatic response. Moreover, the inference from the behavior to the belief seems more reliable than the inferences from the somatic responses to the beliefs.

The fourth set of experiments, on patients suffering from anosognosia, in this case paralyzed patients who believed that they are not paralyzed, provide the most compelling scientific evidence for the existence of self-deception. Specifically, the patients suffer from paralysis of the left side of the body, notably the arm and the hand, caused by a stroke in the right brain hemisphere. One the one hand, these patients firmly believe that they are not paralyzed, as shown not only by what they say, but also by what they do. Thus when given the choice between a well-remunerated task that requires the use of both hands (tying their shoe laces) and a less well-remunerated task that can be performed with one hand (screwing in a light bulb), they invariably choose the former. If, following a failure, they try again ten minutes later, they make the same choice and claim to have succeeded on the first try.

On the other hand, the patients know that they are in fact paralyzed, as shown by several observations. For ill-understood reasons, the anosognosia disappears when the left ear of a patient is irrigated with cold water. In this state, the patient affirms that he is paralyzed and has been so for several days. When the effect of the irrigation wears off and the patient is reminded of this statement, he claims that he had stated having the use of both hands. The patient seems to remember the fruitless effort to his shoelaces, but the access to this memory is blocked until it is unblocked by the cold water. Although these facts do not show that the blocking is *motivated*, other findings point in that direction. Thus a woman who had tried in vain to tie her shoelaces later affirmed that she had done so "with both hands," a detail that a normal person would omit ("the lady doth protest too much, methinks"). Other patients come up with farfetched explanations for their failures to respond to requests to move their left hand. Finally, when the experimenter injects a saline solution in the paralyzed arm and tells the patient, untruthfully, that it is an anesthetic that will paralyze his arm for a few minutes, he answers "No" to the question whether he can move his arm, presumably because a temporary paralysis is much less threatening than a permanent one. Whether these findings can illuminate self-deception in non-damaged persons remains an open question.

The experimental studies illuminate sharply, perhaps too sharply. Real self-deception, outside the laboratory, is a matter of *clair–obscur* rather than of black and white; of half-believing or averting one's gaze rather than of full awareness contrasted with total unawareness. By the nature of the case, a good novelist is better equipped than the experimenter in capturing these fluid or half-crystallized states. I shall briefly discuss a passage from Stendhal, and then at greater length some analyses from *À la recherche du temps perdu*.

In Stendhal's unfinished novel *Lucien Leuwen*, we encounter the phenomenon of self-deceptive ignorance of love. In the early part of the novel, Lucien and a young widow in the garrison town where he is serving, Mme de Chasteller, have come to love each other deeply, yet are uncertain about each other. She fears that he may be no more than a rake, he that she does not really love him. Whenever he makes a clumsy and tentative advance, she sees it as a reason for doubt about his character; she grows haughty, he is made desperate, and, through his desperation, redeems himself for a while. Gradually they grow closer to each other. Lucien writes her a letter; she, after some soul-searching replies in what she believes to be a severe and uncommunicative tone. In an authorial aside, Stendhal comments as follows: "What would be the point of noting that her reply involved a studied attempt at the haughtiest turns of phrase? Three or four times Leuwen was urged to abandon all hope, the very word *hope* was avoided with an infinite adroitness that made Mme de Chasteller very pleased with herself. Alas, without knowing it she was the victim of her Jesuitical education; she deceived herself, in applying badly, and

unawares, the art of deceiving others which she had been taught at the Sacré-Coeur. She *answered*: everything lay in this word, which she preferred to ignore."

Rather than saying that Mme de Chasteller *knew*, at an unconscious level, that the discouraging letter would be read as an encouragement, Stendhal seems to suggest that she *preferred not to explore* the implications of her act. I now discuss some passages where Proust makes a similar suggestion. Because of the conceptual richness and psychological acuity of his analyses, I shall reproduce them at some length.

The question I shall discuss is whether Swann, a friend of the Narrator's family, deceived himself when he thought that his mistress Odette was faithful to him. The stage is set in a passage where Swann reflects on the reasons Odette might have for staying with him, apart from an uncertain and fragile personal attraction:

> For the moment, by lavishing presents upon her and performing all manner of services, he could rely on advantages not contained in his person or in his intellect, and forego the exhausting effort to please by himself. And the price he paid . . . for this delight in being in love, in living by love alone, a delight in whose reality he sometimes doubted, *enhanced its value in his eyes* – as one sees people who are doubtful whether the sight of the sea and the sound of its waves are really enjoyable, become convinced that they are, as also of the rare quality of their own disinterested taste, when they have agreed to pay a hundred francs a day for a room in an hotel, from which that sight and that sound may be enjoyed.

Pursuing this train of thought, Swann reverses the argument:

> One day, when reflections of this order had brought him once again to the memory of the time when someone had spoken to him of Odette as of a "kept" woman, and when, once again, he had amused himself with contrasting that strange personification, the "kept" woman . . . with that Odette upon whose face he had watched the passage of the same expressions of pity for a sufferer, indignation of an act of injustice, gratitude for an act of kindness, which he had seen, in earlier days, on his own mother's face, and on the faces of friends; that Odette, whose conversation had so frequently turned on the things that he himself knew better than anyone, his collections, his room, his old servant, his banker, who kept all his title-deeds and bonds; – the thought of the banker reminded him that he must call on him shortly, to draw some money. And indeed, if, during the current month, he were to come less liberally to the aid of Odette in her financial difficulties than in the month before, when he had given her five thousand francs, *if he refrained from offering her a diamond necklace for which she longed, he would be allowing her admiration for his generosity to decline, that gratitude which had made him so happy, and would even be running the risk of her imagining that his love for her (as she saw its visible manifestations grow fewer) had itself diminished.*[15]

[15] Comparing the phrases I have italicized in these two passages, we note a complete inversion. First, Swann's generosity is explained by its magical capacity to *increase Odette's value in his*

Then, self-deception makes an appearance:

> And then, suddenly, he asked himself whether that was not precisely what was implied by "keeping" a woman (as if, in fact, that idea of "keeping" could be derived from elements not at all mysterious nor perverse, but belonging to the intimate routine of his daily life, such as that thousand-franc note, a familiar and domestic object, torn in places and mended with gummed paper, which his valet, after paying the household accounts and the rent, had locked up in a drawer in the old writing-desk whence he had extracted it to send it, with four others, to Odette) and whether it was not possible to apply to Odette, since he had known her (for he never imagined for a moment that she could ever have taken a penny from anyone else, before), that title, which he had believed so wholly inapplicable to her, of "kept" woman. *He could not explore the idea further*, for a sudden access of that mental lethargy which was, with him, congenital, intermittent and providential, happened, at that moment, to extinguish every particle of light in his brain, as instantaneously as, at a later period, when electric lighting had been everywhere installed, it became possible, merely by fingering a switch, to cut off all the supply of light from a house. His mind fumbled, for a moment, in the darkness, he took off his spectacles, wiped the glasses, passed his hands over his eyes, but saw no light until he found himself face to face with a wholly different idea, the realization that he must endeavor, in the coming month, to send Odette six or seven thousand-franc notes instead of five, simply as a surprise for her and to give her pleasure.

In a sense, Swann "believes" that Odette is a kept woman, and in a sense he is motivated to suppress that belief.[16] It would be misleading to say, however, that the belief is relegated to the unconscious. Like the sun, the belief is too hard to look in the face. As a French historian said about his adherence to the Communist Party in the 1950s, he and his friends "knew what the Soviet Union was like; so, to remain Communists, we made an effort *not to think about it*" (his italics).

### Motivated framing

These are examples of motivated blindness. I conclude by considering what one might call *motivated framing*. People may be opposed to an action or a policy if it is framed in one way, but accept it if it is brought under a different heading. The classical discussion of the issue occurs in Pascal's *Provinciales*, perhaps the greatest satirical work ever written and a blow from which its target, the Jesuit order, never fully recovered. The focus of his criticism is precisely the Jesuitical idea of making forbidden actions appear as licit by *directing one's intention* in the appropriate way. The Jesuits claimed, for

*eyes*. Next, Swann makes the more down-to-earth observation that the generosity might *increase his value in her eyes*.

[16] The word "providential" is puzzling, since the extinction of Swann's mental lights is surely motivated rather than accidental. Other passages from the novel confirm this interpretation.

instance, that even though the Bible prohibits revenge, "a military man may demand satisfaction on the spot from the person who has injured him – not, indeed, with the intention of rendering evil for evil, but with that of preserving his honor." In other words, when A suffers an injury at the hands of B, A may licitly inflict an injury on B by directing his intention to the fact that he would be blamed by C, D, E . . . if he failed to retaliate.[17] In this case as well as in the other examples Pascal considers, the framing is done by the Jesuit confessor, not by the agent himself. I now consider cases in which the agents themselves were responsible for the framing.

In the making of the American constitution of 1789 and the French constitution of 1791, the issue of slavery was prominent in several ways. Yet in both constituent assemblies, the framers took great care to avoid the words *slave* or *slavery* in the constitution they adopted. In America, the framers were simply hypocritical. In their internal deliberations at the Federal Convention, they did not shy away from using the terms, which occur more than a hundred times in Madison's highly compressed notes. Yet to accommodate Northern and perhaps international audiences, they substituted euphemisms such as "other persons" in the text of the document. In France, Robespierre's near-hysterical insistence on using the term "unfree persons" about the slaves in Santo Domingo suggests that, for him, the framing was all-important. He was willing, he said, to let the colonies perish rather than having the word "slave" appear in the law, since, if it did, the declaration of the rights of man and of liberty would be undermined.

These two cases, especially the first, are closer to deception than to self-deception. A case closer to the latter end of the continuum is provided by the way the Soviet elite under Stalin framed their privileges. A scholar of the period notes that since according to Marxist doctrine class societies were based on property, "the fact that the amenities of life – car, apartment, dacha – were not owned but were state issue was very important in enabling Communists of the *nomenklatura* to see themselves as something different from a new nobility or ruling class. On the contrary, they were people who owned nothing! . . . It was comparatively easy for elite members to see themselves as indifferent to material things when there was no personal property at stake."

As a further example, consider the framing of subsidies and taxes. The aluminum industry in western Norway demands and gets huge subsidies in the form of cheap energy, partly because the workers do not want wage subsidies.

[17] The claim that one may legitimately use a bad *means* (killing an adversary) to achieve a good end (defending one's honor) must be distinguished from the doctrine of double effect, according to which it is permissible to bring about as a foreseeable *side effect* of action what it would not be permissible to bring about as the intended result of action. As philosophers have noted, that doctrine also creates a potential for self-deception.

The government has tried to offer the fishermen in northern Norway direct labor subsidies, only to be met with the response that they prefer subsidies to be given to the shipowners. In both cases, observers emphasize that accepting wage subsidies is perceived to be like begging. Workers in the textile industry, where direct wage subsidies *are* given, envy the aluminum industry its energy requirements, because these justify less transparent income transfers. For similar reasons, farmers in western Germany favor a subsidy through price over a direct subsidy of income although this involves astronomical dead-weight losses. Somehow, what would be rejected as a one-step operation becomes acceptable as a two-step operation. Similarly, pacifists may be willing to pay taxes that are used for military purposes. In the War of 1812, the American "government exempted religious pacifists – Quakers, Mennonites, and Dunkers – from the militia, but they had to pay annual fines of five pounds per man and had to provide draft animals, wagons, cars and sleighs on military demand. The Mennonites and Dunkers accepted that compromise, but the Quakers balked at contributing anything that promoted bloodshed."

As the examples of the two previous paragraphs show, social agents can be motivated to frame their situation in a way that is consistent with their preferred self-image as non-exploitative, non-parasitical or non-murderous. As in many other cases of self-deception, the alternative frame is not relegated to the unconscious, but simply not activated or explored. For a more adequate analysis, one would need a Proust.

### Bibliographical note

Evidence for many of the findings reported here can be found in the following source books: D. Kahneman, P. Slovic, and A. Tversky (eds.), *Judgment Under Uncertainty* (Cambridge University Press, 1982); D. Bell, H. Raiffa, and A. Tversky (eds.), *Decision Making* (Cambridge University Press, 1988); T. Connolly, H. Arkes, and K. R. Hammond (eds.), *Judgment and Decision Making* (Cambridge University Press, 2000); D. Kahneman and A. Tversky (eds.), *Choices, Values, and Frames* (Cambridge University Press, 2000); T. Gilovich, D. Griffin, and D. Kahneman (eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment* (Cambridge University Press, 2002); C. Camerer, G. Loewenstein, and M. Rabin (eds.), *Advances in Behavioral Economics* (New York: Russell Sage, 2004); I. Brocas and J. Carillo (eds.), *The Psychology of Economic Decisions*, vols. I and II (Oxford University Press, 2003, 2004). The quote on the veil of ignorance in Rome is from H. Scullard, *From the Gracchi to Nero* (London: Routledge, 1982), p. 35. The comments on the "Texas sharpshooter effect" and spurious cancer statistics are from G. Johnson, *The Cancer Chronicles* (New York: Knopf, 2013), and A. Gawande, "The cancer-cluster myth," *The New Yorker,* February 8,

1999. The observation about the Londoners' perception of bombing patterns is from W. Feller, *An Introduction to Probability Theory and its Applications* (New York: Wiley, 1968), p. 160, and the comment on investors who rely on past performance of firms is from N. Taleb, *Fooled by Randomness* (New York: Random House, 2005). A study of magical thinking in actions that "tempt fate" is A. Arad, "Avoiding greedy behavior in situations of uncertainty: the role of magical thinking," *Journal of Behavioral and Experimental Economics* 53 (2014), 17–23. The cold-water study is in G. Quattrone and A. Tversky, "Causal versus diagnostic contingencies: on self-deception and the voter's illusion," *Journal of Personality and Social Psychology* 46 (1984), 237–48. The greater tendency to impute cooperation to partners than to non-partners is documented in L. Messé and J. Sivacek," Predictions of others' responses in a mixed-motive game: self-justification or false consensus?" *Journal of Personality and Social Psychology* 37 (1979), 602–7. The double incompetence of the ignorant is documented in J. Kruger and D. Dunning, "Unskilled and unaware of it," *Journal of Personality and Social Psychology* 77 (1999), 1121–34. A sophisticated study of the unreliability of the judgments of some experts ("hedgehogs") and the somewhat more reliable judgments of others ("foxes") is P. Tetlock, *Expert Political Judgment* (Princeton University Press, 2005). For Abraham Wald's insight, see J. Ellenberg, *How Not to be Wrong* (New York: Penguin, 2014), pp. 5–7. The example of the man who was puzzled about the railway company's knowledge about his location is lifted from D. Sand, *The Improbability Principle* (New York: Scientific American, 2014), p. 122. For the data about earthquakes and floods, see P. Slovic, *The Perception of Risk* (Sterling, VA: Earthscan, 2000). For the (il)logic of conspiracy theories, see B. Keeley, "Of conspiracy theories," *Journal of Philosophy* 96 (1999), 109–26. On theories of famine see S. Kaplan, "The famine plot persuasion in eighteenth-century France," *Transactions of the American Philosophical Society* 72 (1982), and F. Ploux, *De bouche à oreille: naissance et propagation des rumeurs dans la France du XIXe siècle* (Paris: Aubier, 2003). A study of conspiratorial thinking is R. Hofstadter, *The Paranoid Style in American Politics* (Cambridge, MA: Harvard University Press, 1964). The study of the constraints on motivated reasoning is W. Klein and Z. Kunda, "Motivated person perception: Constructing justifications for desired beliefs," *Journal of Experimental Social Psychology* 28 (1992), 145–68. The best overview of self-deception is a special issue of *Behavioral and Brain Sciences* 20 (1997), organized around an article by A. Mele, "Real self-deception." The "canonical definition" is that of R. Gur and H. Sackeim, "Self-deception: A concept in search of a phenomenon," *Journal of Personality and Social Psychology* 37 (1979), 147–69. The study of homophobia and homosexuality is H. Adams, L. Wright, and B. Lohr, "Is homophobia associated with homosexual arousal?" *Journal of Abnormal Psychology* 105 (1996), 440–5. The

findings about anosognosia are from V. S. Ramachandran and S. Blakelee, *Phantoms in the Brain* (New York: William Morrow, 1999). I discuss Proust's analyses of self-deception at greater length in Chapter 15 of *L'irrationalité* (Paris: Seuil, 2010). The quote from the French historian is taken from P. Veyne, *Et dans l'éternité je ne m'ennuierai pas* (Paris: Albin Michel, 2014), p. 93. For the verbal denials of slavery in the American and French constitutions, see my "Throwing a veil over inequality," in C. Sypnovich (ed.), *The Egalitarian Conscience: Essays in Honour of G. A. Cohen* (Oxford University Press, 2006). On the framing of privilege by the Soviet elites, see S. Fitzpatrick, *Everyday Stalinism* (University of Chicago Press, 1999), pp. 104–5. For references to the framing of subsidies to Norwegian and German workers, see my *Alchemies of the Mind* (Cambridge University Press, 1999), pp. 252–3. The payment of taxes to finance the 1812 War is cited from A. Taylor, *The Civil War of 1812* (New York: Knopf, 2010), p. 310.

# 8 Emotions

### The role of the emotions

Emotions enter human life in three ways. At their most intense they are the most important *sources of happiness and misery*, far overshadowing hedonic pleasures and physical pain. The radiant love of Anne Elliott at the end of *Persuasion* is unsurpassable happiness. A more reproducible if less exalted effect is that of watching Fred Astaire dancing. Conversely, the emotion of shame can be utterly devastating. Voltaire wrote, "To be an object of contempt to those with whom one lives is a thing that none has ever been, or ever will be, able to endure."

Shame also illustrates the second way in which emotions matter, namely, in their *impact on behavior*. In Chapter 4, I cited several cases in which people killed themselves because of the overpowering emotion of shame. In this chapter I shall mainly discuss the *action tendencies* that are associated with the emotions. The extent to which these tendencies are translated into actual behavior will concern us in later chapters.

Third, emotions can matter because of their impact on *other mental states* – on motivations as well as on beliefs. When a desire for a certain state to obtain is supported by a strong emotion, the tendency to believe that it does obtain can be irresistible. As Stendhal says in *On Love*, "From the moment he falls in love even the wisest man no longer sees anything *as it really is* . . . He no longer admits an element of chance in things and loses his sense of the probable; judging by its effect on his happiness, whatever he imagines becomes reality." In *À la recherche du temps perdu* Proust pursues the same theme over hundreds of pages, with more variations and twists than one might have thought possible.

Also, when an emotion triggers a negative meta-emotion, such as shame, the pressure to transmute it into a more acceptable emotion may be irresistible (see next chapter). This mechanism presupposes, of course, that the agent is aware of the emotion and identifies it correctly. Anger and love sometime creep up on us, without our being aware of them until they suddenly erupt into consciousness. In *Emma*, Jane Austen offers a hilarious description of a young woman

who mistakes her state of boredom for love. "This sensation of listlessness, weariness, stupidity, this disinclination to sit down and employ myself, this feeling of every thing's being dull and insipid about the house! – I must be in love."

## What are the emotions?

Before considering each of these aspects of emotion in more detail, I need to say something about *what emotions are* and *what emotions there are*. There is no agreed-upon definition of what counts as an emotion, that is, no agreed-upon list of sufficient and necessary conditions. There is not even an agreed-upon list of necessary conditions. Although I shall discuss a large number of common features of the states that we understand, preanalytically, as emotions, there are counterexamples to all of them. For any such feature, that is, there are some emotions or emotional occurrences in which it is lacking. We may think that action tendencies are crucial to emotion, but the aesthetic emotions provide a counterexample. We may think that a "short half-life," that is, a tendency to decay quickly, is an essential feature of emotion, but in some instances unrequited romantic love (such as that of Cyrano de Bergerac) or the passionate desire for revenge can persist for years or decades.[1] We may think that emotions are triggered by beliefs, but how do we then explain that people can get emotionally upset by reading stories or watching movies that are clearly fictitious? Many other examples could be given of allegedly universal features that turn out to be lacking in some cases.

In light of this problem, the natural response is to deny that "emotion" is a useful scientific category. In the language of philosophers, emotions do not seem to form a *natural kind*. In spite of their difference, whales and bats, qua mammals, belong to the same natural kind. Whales and sharks, in spite of their similarity, do not; nor do bats and birds. Anger and love have in common the capacity for clouding and biasing the mind, but this similarity does not make them into a natural kind. To see how such reasoning by analogy can go astray, we may notice that the intake of amphetamines and romantic love produce many of the same effects: acute awareness, heightened energy, reduced need for sleep and food, and feelings of euphoria. In fact, the states of hypomania, intense creative activity, and enthusiasm partake of the same features. Yet nobody would claim, I assume, that these five states belong to the same natural kind.[2]

[1] Thus Seneca was wrong, for once, when he asserted that if a delay makes no difference for the intention to act, we can infer that the agent was not under the sway of anger.

[2] They may, however, recruit some of the same neural circuitry.

For the purpose of social-scientific explanation, this conundrum can be left unresolved. We can focus on occurrences of emotions in which a certain number of features are regularly observed and ask how these can help us to explain behavior or other mental states. The fact that in other occurrences that intuitively count as emotions some of these features are lacking is interesting from a conceptual point of view but does not detract from their explanatory efficacy in cases where they are present. I shall list and comment on the features to which I want to draw attention:

- *Cognitive antecedents*. Most emotions are triggered by beliefs, often by the agent's acquisition of a new belief. Emotions may also have other causal conditions (we are more readily irritated when we are tired), but the presence of these will not by themselves cause the emotion to occur, any more than a slippery road will cause a car accident. As explained later, these beliefs may fall short of full certainty.
- *Perceptual antecedents*. Some emotions, such as disgust and fear, can be triggered by mere perceptions, prior to belief formation. Emotions that are triggered by cognition may be strengthened if there is also a perceptual element present, as when the fear triggered by warnings on cigarette packs is enhanced by graphic color pictures of cancer lungs.[3]

Neurophysiological work on fear (in rats) confirms this idea. There are two different pathways from the sensory apparatus in the thalamus to the amygdala (the part of the brain that causes visceral as well as behavioral emotional responses). Confirming the traditional view that emotions are always preceded and triggered by a cognition, one pathway goes from the thalamus to the neocortex, the thinking part of the brain, and from the neocortex onward to the amygdala. The organism receives a signal, forms a belief about what it means, and then reacts emotionally. There is also, however, a direct pathway from the thalamus to the amygdala that bypasses the thinking part of the brain entirely. Compared to the first pathway, the second is "quick and dirty." On the one hand, it is faster. In a rat it takes about twelve milliseconds (twelve one-thousandths of a second) for an acoustic stimulus to reach the amygdala through the thalamic pathway, and almost twice as long through the cortical pathway. On the other hand, the second pathway differentiates less finely among incoming signals. Whereas the cortex can figure out that a slender curved shape on a path through the wood is a curved stick rather than a snake, the amygdala cannot make this distinction. Yet from the point of view of survival, the cost of reacting to a stick as if it were a snake must have been much smaller than the cost of the opposite mistake.

[3] Research indicates that the larger graphic Canadian health warnings has a greater impact than the smaller US text warnings and the small text warnings in Mexico. Similarly, graphic warnings in Thailand have greater impact than text-based Malaysian warnings.

I do not know whether these findings from the study of fear generalize to other emotions. Conjecturally, something of the sort might also be true of anger. When exposed to something that could be an attack, the opportunity cost of waiting to find out whether it is one might be very high. Natural selection might well have hardwired a tendency to "shoot first; ask later." If I do lash out and later find out that I was in fact not the victim of an attack, I might nevertheless invent a story to justify my behavior. This rather subtle mechanism, linked to our amour-propre (Chapter 9), would interact with a neurophysiological mechanism that we share with animals that lack the need for self-esteem. This is likely to be the pattern of many findings from physiology and neuroscience. Near-automatic reactions that we share with other species may be subject to the self-serving interpretations and elaborations that are unique to human beings. These rationalizations are not trivial, since they may cause us to persist in aggression rather than admit that we were at fault.

The importance of perception in generating emotions can be illustrated by the attitude of White House officials toward the loss of American soldiers in Vietnam. In 1964, McGeorge Bundy and his staff compared casualties in Vietnam to traffic-related injuries in Washington DC and found them insignificant. In early 1965, the Chairman of the Joint Chiefs of Staff, Maxwell Taylor, recommended that Bundy be sent on a mission to Vietnam because of "the fact that he has been physically detached from the local scene and hence would have an objectivity which an old Vietnam hand would lack." When he witnessed the effects of an attack at Pleiku, in the central highlands of South Vietnam, Bundy lost his detachment. The "sight of gravely wounded American servicemen evoked in Bundy uncharacteristically strong and visible emotion."

- *Physiological arousal.* Emotions go together with changes in heart rate, electrical skin conductance, bodily temperature, blood pressure, respiration, and numerous other variables.
- *Physiological expressions*. Emotions go together with characteristic observable signs, such as bodily posture, voice pitch, flushing and reddening (from embarrassment), smiling or baring the teeth, laughing and frowning, weeping and crying, and white or red anger (as manifested in pallor and blushing, respectively).
- *Action tendencies*. Emotions are accompanied by tendencies or urges to perform specific actions. Although these tendencies may not lead to actual behavior, they are more than dispositions – they are forms of incipient behavior rather than mere potential for behavior. I discuss some of these tendencies at greater length later.
- *Intentional objects*. Unlike other visceral phenomena such as pain or hunger, emotions are *about* something. They may have "propositional

objects" ("I am indignant that . . .") or non-propositional objects ("I am indignant with . . ."). As we shall see, the lack of a clearly defined object may prevent the emotion from being triggered.

- *Valence*. This is a technical term for the pain-pleasure dimension of the emotions as we experience them. As noted, the valence might range from the glowing happiness of Anne Elliott to the crushing shame of the exposed consumers of pedophiliac material.

Do not emotions, like colors, also have specific qualitative *feelings*? Shame and guilt, for example, seem to *feel* different in a way that cannot be reduced to the fact that shame is more intensely unpleasant. There is evidence that one could insert an electrode into my brain and make me feel sad, embarrassed, or afraid even though I would not be able to identify either a cause or an object of the feeling. Important as this aspect may turn out to be for our understanding of emotion, it is not yet well enough understood to suggest specific causal hypotheses.

### What emotions are there?

I shall list and briefly describe some two dozen emotions, without claiming that this classification is superior to the many others that have been proposed. My aim is to provide some understanding of the emotions that have either intrinsic or causal importance in social life, not to try to satisfy the (legitimate) concerns of emotion theorists. In particular, I shall have nothing to say about which emotions are "basic" or "non-basic."

One important group of emotions are the *evaluative emotions*. They involve a positive or a negative assessment of one's own or someone else's behavior or character.[4] If an emotion is triggered by the behavior of another person, that behavior may be directed either toward oneself or toward a third party. These distinctions yield ten (or eleven) emotions altogether:

- *Shame* is triggered by another person's negative belief about the agent's character.
- *Contempt* and *hatred* are triggered by the agent's negative beliefs about another's character. Contempt is induced by the thought that another is inferior, hatred by the thought that he is evil. Hitler thought Jews were evil, and Slavs inferior.
- *Guilt* is triggered by a negative belief about one's own action.

[4] Emotions triggered by negative assessments of oneself always have negative valence. Those caused by negative assessments of others are more ambiguous in this respect. Some people enjoy being angry or indignant, and even seem to seek out occasion to trigger these emotions. If they do, the mechanism is probably reinforcement (Chapter 11) rather than intentional choice.

- *Anger* is triggered by a negative belief about another's action toward oneself.[5]
- *Cartesian indignation*[6] is triggered by a negative belief about another's action toward a third party.
- *Pridefulness* is triggered by a positive belief about one's own character.
- *Liking* is triggered by a positive belief about another's character.
- *Pride* is triggered by a positive belief about one's own action.
- *Gratitude* is triggered by a positive belief about another's action toward oneself.
- *Admiration* is triggered by a positive belief about another's action toward a third party.

Second, there is a set of emotions generated by the thought that someone else is in the deserved or undeserved possession of some good or bad.[7] The target of these emotions is neither individual action nor individual character, but a state of affairs. Following Aristotle's discussion in the *Rhetoric*, we may distinguish six (or seven) cases.

- *Envy* is caused by the deserved good of someone else.
- *Aristotelian indignation* is caused by the undeserved good of someone else.[8] The closely related emotion of *resentment* is caused by the reversal of a prestige hierarchy, when a formerly inferior group or individual emerges as dominant.
- *Sympathy* is caused by the deserved good of someone else.
- *Pity* is caused by the undeserved bad of someone else.
- *Malice* is caused by the undeserved bad of someone else.
- *Gloating* is caused by the deserved bad of someone else.

Third, there are positive or negative emotions generated by the thought of good or bad things that have happened or will happen – *joy* and *grief*, with their several varieties and cognates. The emotion of *enthusiasm* – neglected by emotion theorists but observed in many revolutionary moments – also belongs

[5] Anger can also be triggered by the sheer frustration of a desire. Seneca cites the example of the Romans who showed their anger at gladiators when they were unwilling to die. Closer to us, many drivers are angry at cyclists who force them to slow down. The anger can induce a rewriting of the script (Chapter 9), to make drivers believe that the cyclists are deliberately slowing them down. If the latter are subject to reactance and refuse to be crowded (Chapter 9), the belief may actually be justified.

[6] The emotion was first identified by Descartes, who added the important qualification that when the agent *loves* the third party the reaction is anger rather than indignation.

[7] I include "non-undeserved" under the heading of "deserved." Thus when someone wins the big prize in the lottery I shall say that it is deserved, contrary to ordinary usage.

[8] Although Aristotle's term for this emotion is usually translated by "indignation," it should be clear how it differs from Cartesian indignation.

here. As many have observed, bad events in the past may also generate positive emotions in the present, and good events negative emotions. Thus in the main collection of proverbial sayings from antiquity, the *Sentences* of Publilius Syrus, we find both "The remembrance of past perils is pleasant" and "Past happiness augments present misery."

All the emotions discussed so far are induced by beliefs that are (or may be) held in the mode of certainty. There are also emotions – *hope*, *fear*, *love*, and *jealousy* – that essentially involve beliefs held in the modes of probability or possibility. These emotions are generated by the thought of good or bad things that may or may not happen in the future, and of good or bad states of affairs that may or may not obtain in the present. By and large, these emotions require that the event or state in question be seen as more than merely conceivable; that is, there must be a non-negligible chance or a "downhill causal story" that it might actually occur or obtain. The thought of winning the big prize in the lottery may generate hope, but not the "uphill" thought of receiving a large gift out of the blue from an unknown millionaire. These emotions also seem to require that the event or state fall short of being thought to be certain. If I *know* that I am about to be executed, I may feel despair rather than fear. According to Stendhal, love withers away both when one is certain that it is reciprocated and when one is certain that it is not. According to La Rochefoucauld, jealousy may disappear the moment one *knows* that the person one loves is in love with somebody else.

Some emotions are generated by *counterfactual* thoughts about what might have happened or what one might have done. *Disappointment* is the emotion that occurs when a hoped-for positive event fails to materialize.[9] *Regret* is the emotion that occurs when we realize we could have made a hoped-for positive event occur if we had made a different choice. The positive counterparts of these emotions (caused by the non-occurrence of negative events) are sometimes referred to as *elation* and *rejoicing*, respectively. (In everyday language, the two are usually blurred under the heading of *relief*.) Whereas disappointment and elation involve comparisons of different outcomes caused by different states of the world for a given choice, regret and rejoicing involve comparisons caused by different choices within a single state. In some cases, negative events can be imputed to either source. If I get wet on my way to work I may either ascribe it to a chance meteorological event or to the fact that I did not take an umbrella. Although I might prefer the first framing, this piece of wishful thinking might be subject to reality constraints (Chapter 7) if I had just heard a forecast of rain before leaving the house.

[9] Near-misses generate stronger emotions; thus silver medalists in Olympic competitions report less happiness than do bronze medalists.

Some emotions, finally, are generated by other emotions. I shall have more to say about such meta-emotions in Chapter 9. Here I shall only mention shame at feeling envy or guilt at loving an inappropriate person.

### Emotions and happiness

The role of emotions in generating happiness (or misery) suggests the idea of a "gross national happiness product." The usual measures of economic performance are, of course, more objective. Yet objectivity in the sense of physical measurability is not what we ultimately care about. The reason we want to know about economic output is that it contributes to *subjective* welfare or happiness. Moreover, happiness can stem from sources that do not lend themselves to any kind of objective quantitative measurement. In 1994, when Norway hosted the Olympic Winter Games, the country had to build new arenas for the events and housing for the participants, at considerable cost. On the revenue side one could include the money spent by visitors to the country and by spectators of the events, as well as the income generated by these constructions in the future. Economists who have carried out these calculations do not believe that the games broke even. I feel utterly certain, however (but of course cannot prove), that if we include the emotional benefits to the Norwegian population, the games ran a huge surplus. The unexpectedly large number of Norwegian gold medalists created a mood of collective euphoria, which was all the greater *because* the victories were so unexpected. The "objective" number of victories owed a great deal of its impact to the element of subjective surprise.[10] More recently, the victory of the French soccer team in the 1998 World Cup as well as its defeat in 2002 generated feelings of euphoria and dejectedness that owed much of their intensity to the fact of surprise.

In general, it is difficult to compare the emotional components of welfare or well-being with other components. That positive emotions at their most intense contribute more to happiness than does simple hedonic welfare proves nothing, unless we know how often the intense episodes occur. Also, we do not understand whether and to what extent the propensity for emotional highs goes together with the propensity for emotional lows. If it does, is a life in steady contentment happier overall than one that alternates between euphoria and dysphoria? As Montaigne noted, the answer depends on the occasions offered by the environment. "If you say that the convenience of having our senses chilled and blunted when tasting evil pains must entail the

[10] Suppose that the prior probability of a victory is p and that the satisfaction derived from a victory is proportional to 1/p (because the surprise is greater when p is low). In this special model of surprise, the *expected* satisfaction of victory is independent of the probability of victory. In general, the impact of surprise is going to be more complex.

consequential inconvenience of rendering us less keenly appreciative of the joys of good pleasures, I agree. But the wretchedness of our human condition means that we have less to relish than to banish." The ideal of extinguishing the emotions that one finds in many ancient philosophies, notably Stoicism and Buddhism, emerged in societies where the environment may have offered more occasions for emotions with negative valence. Writing during the wars of religion that were devastating France, Montaigne may have been in the same situation.

### Emotion and action

The mediating link between emotion and action is that of an action tendency (or action readiness). We may also think of an action tendency as a temporary preference. Each of the major emotions seems to have associated with it one (or a few) such tendencies (see Table 8.1).

Although anger and Cartesian indignation induce the same action tendency, that of anger is stronger. Experiments show that subjects are willing to incur a larger cost to punish somebody who hurt them than to punish somebody who hurt a third party. After the end of World War II, Americans were often more eager to punish Nazis who had mistreated American prisoners of war than those who were responsible for the Holocaust. An exception, which confirms the principle, were the Jewish members of the Roosevelt administration.

The emotions of anger, guilt, contempt, and shame have close relations to moral and social norms. Norm violators may suffer guilt or shame, whereas those who observe the violation feel anger or contempt. The structures of these relations differ as shown in Figure 8.1. Social norms, further discussed in Chapter 21, are mediated by exposure to others. That is why the suicides mentioned in Chapter 5 occurred only when the shameful actions

Table 8.1

| Emotion | Action tendency |
|---|---|
| Anger or Cartesian indignation | Cause the object of the emotion to suffer |
| Hatred | Cause the object of hatred to cease to exist |
| Contempt | Ostracize; avoid |
| Shame | "Sink through the floor"; run away; commit suicide |
| Guilt | Confess; make repairs; hurt oneself |
| Envy | Destroy the envied object or its possessor |
| Fear | Flight; fight |
| Love | Approach and touch the other; help the other; please the other |
| Pity | Console or alleviate the distress of the other |
| Gratitude | Help the other |

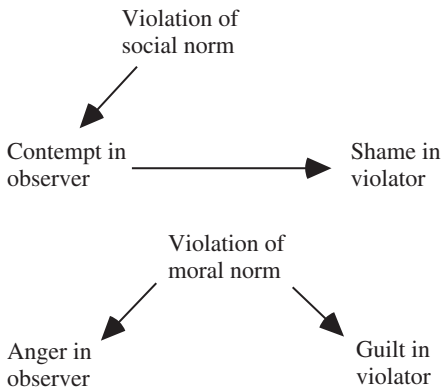



Figure 8.1

became public knowledge. As I argued in Chapter 5, moral norms differ in this respect.[11]

Some action tendencies appear to aim at "restoring the moral balance of the universe." Hurting those who hurt you and helping those who help you are, seemingly, ways of *getting even*. This may be true in some cases. The moral-balance view is compelling in the case of guilt, where the action tendency of making repairs is explicitly restorative. Moreover, when the agent cannot undo the harm she has done, she can restore the balance by harming herself to an equal extent. If I have cheated on my income taxes and discover that the IRS does not accept an anonymous money order for the amount I owe, I can restore the balance by burning the money instead. As Montaigne noted, the act of repentance has to *cost* the agent something: "I have seen many in my time smitten in conscience for having withheld other men's goods who arrange in their testament to put things right after they are dead. But it is valueless to fix a date for so urgent a matter or to wish to right wrongs without feeling or cost." Earlier, I cited Oscar Wilde's observation to the same effect.

The action tendency of anger, to take revenge, is more complex. In particular, it is probably not expressed by the maxim "an eye for an eye." The *Lex talionis* served to *limit* the extent of revenge rather than to create an obligation to take revenge. It forbids the taking of two eyes for one or an eye for a tooth. The Koran, too, says that "If you want to take revenge, the action should not exceed the offense" (Sura XVI). In this perspective, the *Lex talionis* would serve to counteract a spontaneous tendency to excessive revenge. That

[11] I suspect that violations of quasi-moral norms trigger the same emotions as do violations of moral norms, but my intuition is not robust.

tendency may, instead, be "*Two* eyes for an eye."[12] A character in one of Seneca's plays asserted that "a wrong not exceeded is not revenged." In response to the killing of 365 Lebanese Muslims a Lebanese woman said that "at this moment I want the [Moslem militia] . . . to go into offices and kill the first seven hundred and thirty defenseless Christians they can lay their hands on." In Greece in 1944, after an attack by left-wing groups on the right-wing leader Nikolaos Papageorgiou, "two young men were found dead in the road where the original attack had taken place; placards around their necks read 'Papageorgiou – 2 for 1.'"[13]

The asymmetry might also be due to loss aversion (see Chapter 13), that is, the tendency to value losses roughly twice as heavily as gains. In the following passage, I have transposed an application of the theory of loss aversion to negotiated disarmament to the case of revenge. Italicized statements in parenthesis indicate the transpositions.

> Loss aversion, we argue, could have a significant impact on conflict resolution. Imagine two countries negotiating the number of missiles that they will keep and aim at each other (*two clans involved in a long-standing conflict*). Each country derives security from its own missiles and is threatened by those of the other side. (*Each clan finds security in numbers and is threatened by numbers on the other side.)* Thus, missiles (*persons*) eliminated by (*on*) the other side are evaluated as gains, and missiles (*persons*) one must give up (*killed on one's own side*) are evaluated as losses, relative to the status quo. If losses have twice the impact of gains, then each side will require its opponent to eliminate (*lose*) twice as many missiles (*persons*) as it eliminates (*loses*).

To further complicate matters, excessive retaliation might aim at *deterrence* rather than at revenge. In a draft of *The Civil War in France,* Marx cites an official statement by the Commune that "Every day the banditti of Versailles slaughter or shoot our prisoners, and every hour we learn that another murder has been committed . . . The people, even in its anger, detests bloodshed, as it detests civil war, but it is its duty to protect itself against the savage attempts of its enemies, and whatever it may cost, it shall be an eye for an eye, a tooth for a tooth." He also cites, however, a letter from a priest taken hostage by the

[12] There is some experimental evidence for this mechanism. When instructed to apply the same force on another participant as the latter had applied to them, subjects escalated on average 38 percent in each round. The explanation is that the perception of force is attenuated when the force is self-generated. Although this mechanism may explain escalation in fighting among children, it would not apply to less direct forms of interaction. In some trust games with punishment (see Chapter 20), the punishment that investors imposed on the trustee who did not return part of the profit on their investment was substantially larger than the loss he inflicted on them.

[13] Both Seneca and Adam Smith affirm that a similar asymmetry exists in *gratitude*. You must return more than you received, since returning the equivalent would just be like repaying a loan. Actually, however, because of loss aversion (Chapter 14), repaying the equivalent would be *experienced* as returning *more* than you received.

*communards* which asserts that "a decision has been taken to execute two of the numerous hostages they hold for every new execution" by the government forces. In German-occupied countries during World War II, the ratio was often much higher, up to a hundred to one. These reprisals were often very effective. In actual situations, revenge, loss aversion, and deterrence may interact and reinforce each other to produce a "two for one" or "many for one" pattern, making it hard to identify the pure action tendency.

Emotional action tendencies do not merely induce a desire to act. They also induce a desire *to act sooner rather than later*. To put this idea in context, let me distinguish between *impatience* and *urgency*. I define impatience as a preference for early reward over later reward, that is, some degree of time discounting. As I noted in Chapter 6, emotions may cause an agent to attach less importance to temporally remote consequences of present action. I define urgency, another effect of emotion, as a preference for early action over later action.

To introduce and illustrate the distinction between urgency and impatience, I shall use examples involving money, even though urgency is not likely to be an important factor in monetary decisions. Strictly speaking, therefore, I ought to refer to units of utility rather than to dollars. As the purpose of the examples is merely illustrative, this does not really matter.

In the following statements, all preferences are Monday preferences, i.e. the preferences I hold on Monday:

> *Urgency*: I prefer acting on Monday to get $100 on Wednesday to acting on Tuesday to get $150 on Wednesday.
> *Impatience*: I prefer acting on Monday to get $100 on Tuesday to acting on Monday to get $150 on Wednesday.

Impatience is the familiar difficulty of deferring gratification. There is a two-way trade-off between the size of the reward and the time of delivery of the reward. In urgency, there is a three-way trade-off: the urge to act immediately may be neutralized if the size of the reward from acting later is sufficiently large or if that reward is delivered sufficiently early. In other words, the following statements may all be true:

> I prefer acting on Monday to get $100 on Wednesday to acting on Tuesday to get $150 on Wednesday.
> I prefer acting on Tuesday to get $300 on Wednesday to acting on Monday to get $100 on Wednesday.
> I prefer acting on Tuesday to get $150 on Tuesday to acting on Monday to get $100 on Wednesday.

Since early action and early reward often go together, the choice of the early action may simply be due to the fact that it promises the early reward. In principle, however, it should be possible to tease the two apart.

It is perhaps more intuitive to explain the idea of urgency as a form of *inaction aversion*.[14] In a book on *How Doctors Think*, the author refers to a "tendency toward action rather than inaction. Such an error is more likely to happen with a doctor who is overconfident, whose ego is inflated, but it can also occur *when a physician is desperate and gives in to the urge to 'do something'*. The error, not infrequently, is sparked by pressure from a patient, and it takes considerable effort for a doctor to resist. 'Don't just do something, stand there' ... one of my mentors once said when I was unsure of a diagnosis."

The behavior of terrorists and suicide attackers may also be illuminated by the notion of urgency. Before they kill themselves, suicide attackers are presumably in a state of high emotional tension. According to one Kamikaze pilot, the stress of waiting was unbearable. To counteract the urge to take immediate and premature action, the first rule of the Kamikaze was that they should not be too hasty to die. If they could not select an adequate target, they should return to try again later. In Afghanistan, organizers sometimes prefer the technique of remote detonation, which reduces mistakes caused by attacker stress, such as premature detonation.

The urgency of the emotions provides one of the mechanisms by which they may affect belief formation. As we shall see in Chapter 13, rational belief formation requires an optimal gathering of information. Rather than going by the evidence already at hand, a rational agent will gather additional evidence before acting if the decision to be made is sufficiently important and the cost of waiting sufficiently small. Urgent emotions are often triggered in situations in which the cost of waiting *is* high, that is, in the face of acute physical danger. In such cases, acting quickly without pausing to find out more is of the essence. But when an important decision could be improved by waiting, an emotion-induced desire for immediate action can be harmful. As Seneca said, "Reason grants a hearing to both sides, then seeks to postpone action, even its own, in order that it may gain time to sift out the truth; but anger is precipitate."[15] The proverb "Marry in haste, repent

[14] It is not the only form. Pascal said that "all of humanity's problem stems from man's inability to sit quietly in a room alone." Hume, commenting on a peaceful period in the reign of Henry VII, said that the king and his ministers "might even have dispensed with giving any strict attention to foreign affairs, were it possible for men to enjoy any situation in absolute tranquility or abstain from projects and enterprises, however, fruitless and unnecessary." Keynes wrote that "Most, probably, of our decisions to do something positive, the full consequences of which will be drawn out over many days to come, can only be taken as a result of animal spirits – of a spontaneous urge to action rather than inaction." It is not clear whether these observations refer to the same phenomenon, but they are all distinct from emotional urgency.

[15] The curtailments of civil liberties enacted by several Western governments in the wake of September 11, 2001, are a good test case. Did they illustrate the need to act rapidly in the presence of an imminent danger, or were they a panicky reaction that, by making these

at leisure" suggests both the impetus of emotion and the unfortunate consequences of the inability to resist it.

When the emotion cools down, it may not be possible to undo the actions it triggered. Gibbon writes about the Emperor Theodosius that after his "passion was inflamed" by a minister and he dispatched "the messengers of death, he attempted, when it was too late, to prevent the execution of his orders." In the trials of agents and collaborators of the Germans after 1945, prison sentences were reduced when emotions faded, but executions were irreversible. When young men and women join the Armed Revolutionary Forces of Colombia (FARC) in a moment of enthusiasm and are subsequently disillusioned by the extreme harshness of the organization, they are not allowed to leave it (a "lobster-trap" situation). In 2014, Western recruits to the Islamic State had to hand in their passports when they joined the organization, or their passports were burned.

Urgency is not the only cause of premature belief formation. The need to achieve *cognitive closure* – to form some opinion or other, rather than a specific opinion – can have the same effect. This mechanism, too, is related to emotions, but not in the same way as urgency. Some individuals may be characterized by what Otto Neurath called "an emotional disposition for which the elimination of doubt means a release from a feeling of displeasure." The disposition is also sometimes referred to intolerance of ambiguity or of uncertainty. Whereas urgency is directly induced by the desire to act, the need for cognitive closure may exist even when there is no occasion to act.

Earlier, I mentioned that not all emotions have a short half-life. Usually, nevertheless, emotions do decay quite rapidly with time. In some cases, this is simply due to the fact that the situation that triggered them ceases to exist. When I have gotten safely away from the bear that was threatening me, fear is no longer warranted. More often, though, emotion wanes as memory fades, by the sheer passage of time. Anger, shame, guilt, and love rarely persist with the intensity they had at the onset of emotion. Although McGeorge Bundy's strong emotions at Pleiku caused him to recommend a policy of "sustained reprisal" rather that a punctual tit-for-tat response, "once [his field-marshal] psychosis subsided ... Bundy's belief in the potency of bombing was short-lived." After September 11, 2001, the number of young American men who expressed an interest in serving in the army increased by 50 percent, but there was no marked increase in actual enlistment. These facts are consistent with the hypothesis that the initial surge of interest was due to emotion, which then abated during the several months required for the enrollment process. There

governments appear even more odious in the eyes of their enemies, made further attacks more rather than less probable? Compare also the remarks on the "psychology of tyranny" in earlier chapters.

was almost no increase in the interest in serving among young women, a fact that has no obvious explanation.

This being said, people are often unable to anticipate the decay of their emotions (see later discussion). When in the grip of a strong emotion, they may believe, wrongly, that it will last forever and may even lose any sense of the future. If the suicidal individuals I have referred to had known that their shame (and the contempt of the observers) would abate, they might not have killed themselves. If young couples knew that their love for each other might not last forever, they would be less willing to make binding commitments and in particular to enter into a "covenant marriage" from which it is more difficult to exit.

Let me conclude on this point by pointing to an interaction between two emotion-induced phenomena: preference reversal and clouded belief formation. On a rosy view, these two might cancel one another: because of the preference reversal one wants to act contrary to one's calm and reflective judgment, but because of the clouded belief one is unable to carry out the intention. More frequently, perhaps, the two will reinforce each other. Vengeance is an example. The risk is minimal if I do not take revenge for an affront (disregarding the contempt I might incur from third parties), greater if I take revenge but bide my time; and maximal if I take revenge immediately without any concern for the risks. Montaigne made a similar observation: "When we punish any injuries we have received, philosophy wants us to avoid choler, not so as to diminish our revenge but (on the contrary) so that its blow may be weightier and better aimed; philosophy considers violent emotion to be an impediment to that." Yet that is to ignore the paradox that if we do not feel emotion, we may not want revenge, and if we do feel it, we may not be able to carry out the revenge effectively.

### Emotions and politics

Emotions matter in politics, sometimes in systematic ways. The emotions that affect political decisions and outcomes are mostly negative: anger, Cartesian indignation, hatred, fear, resentment, envy, guilt, and shame. Only one positive emotion, *enthusiasm*, seems capable of affecting political behavior. I shall consider two examples: an important episode in the French Revolution and transitional justice after the end of World War II.

Historians have coined the term "the Great Fear" for the emotion that swept the French countryside in the spring and summer of 1789. In fact, that summer saw not one but two "great fears," the second triggering the decisions made by the deputies in Versailles when they learned about the first. I return to the fear in the countryside in Chapter 22; here I discuss its repercussions on the assembly.

In 1789, "the Great Fear" peaked in late July, leading to numerous "anti-seigneurial actions" such as the sacking of castles, the burning of tenancy records, and occasionally physical attacks on the nobles. When news about these events reached the constituent assembly in early August, many deputies feared for their properties and their families. After an initial urge to crack down on the uprisings, on August 4 the assembly went to the other extreme and abolished feudalism overnight. Many of the deputies gave up their personal privileges or those of their city or province. It is not easy to determine the exact motivational mix behind this "self-denying ordinance," to use a phrase coined about Cromwell's decision in 1645 that members of parliament should exclude themselves from military office.[16] Tocqueville was probably right when he asserted that the decisions were "the combined result, in doses that are impossible to determine, of fear and enthusiasm."[17] Some illustrations and testimonies follow.

Considering the role of *fear*, the emotion is manifest in several letters from August 7 onward by the Comte de Ferrières, a deputy for the nobility, to his wife. The first letter contains very detailed instructions to sell his sheep and his oxen for cash, at any price; to gather all the money and documents in his castle in Mirebeau and transfer them to their house in Poitiers, making sure nobody observes her doing so; to ship their mattresses, bed covers and sheets to Poitiers ("in case of an event, at least something will be saved"). Three days later, he tells her to go with their daughters to Poitiers, even if the harvest should suffer: "do not consider the costs and do not ask to be protected by soldiers, since this would cause alarm in the countryside." He does not care if after these precautions his castle is burned, as he is never going to live there again. His comments on why he is reluctant to cast his votes sincerely in the assembly reflect the same visceral fear.[18]

Considering the role of *enthusiasm*,[19] the *Courrier de Provence*, the organ of the Comte de Mirabeau, referred to "reciprocal challenge and combat in generosity" and to "the seduction of applause, the emulation of outdoing one's

[16] Another such ordinance was the decision of the assembly on May 16, 1791 that its members should be ineligible for the first ordinary legislature. As in the case of the decision on August 4, 1789, the real motivations of many deputies who voted for this measure were very different from the pure disinterestedness (falsely) professed by its proposer, Robespierre.

[17] In his *Recollections,* Tocqueville makes a similar comment on the motivations of the French constituent assembly of 1848 – "fear of outside events and the enthusiasm of the moment."

[18] How do I know that his fear was visceral rater than prudential? The best evidence for my interpretation is the urgency of his reaction and the later reversal of his policy in letters to his wife a few days later, suggesting that an initial emotion had cooled off. Also, the insistent tone of the letters suggests panic.

[19] Apart from some brief and penetrating remarks by Kant, the emotion of enthusiasm has been little studied. While generally praising it, and distinguishing it from sentimentality (*Schwärmerei*), Kant observed that enthusiasm tends to produce good ends but a bad choice of means: the best becomes the enemy of the good.

colleagues, the honor of personal disinterestedness, and the kind of noble intoxication which accompanies the effervescence of generosity." Another contemporary document refers to "the heat of the moment that electrified each individual and made him fear being left behind" in the competition to be generous. To be sure, these statements do not exactly describe a disinterested motivation, but rather an egocentric desire to be *seen* as disinterested. This desire is perhaps an intrinsic or at least inevitable feature of enthusiasm, which tends to arise in crowds and assemblies.[20]

Soon after the capitulation of Germany on May 8, 1945, in some cases even earlier, the countries that had been under German occupation began preparing the prosecution of agents and collaborators of the various regimes. Eventually a proportion of the population ranging from 0.2 percent (Austria) to 2 percent (Norway) suffered some kind of punishment, including the loss of civil and political rights. The number of executions ranged from four per million of population (Austria, Holland) to thirty-nine (France). In most of the countries, the authorities stated that the trials were to be strictly constrained by the rule of law, excluding notably retroactive laws and collective guilt. Justice, not revenge, was the goal. The temporal pattern of the trials suggests, however, that strong emotions were at work. There were clear indications of *urgency*.[21] Maurice Rolland, the official in charge of the early stages of transitional justice in France, asserted that "the government should establish *justice before railroads*." Also, the fact that identical acts of collaboration, such as joining the Waffen SS, received more lenient sentences as time went on strongly suggests that judges and jurors were motivated by emotions with a short-half life.[22] There is also evidence that

[20] A more complex catalogue of the motives in play on August 4, 1789, also offered by a contemporary, is the following: "Some were motivated by the general utility, but many made a virtue of necessity. Some thought they would trap their adversaries, others aspired to praise by newspapers or a group without concern for the consequences; a third was swept up in the general intoxication; and a fourth tried to spoil things by pushing them into extravagance."

[21] In some cases, the urge to act immediately may also be due to an anticipation that emotions will decay over time: "Let's act now, while we are still angry!" This attitude may explain the urgent desire to act of some agents of transitional justice in Belgium in 1944–45, who remembered the slowness of the punishment of collaborators with the Germans after 1918. In general, however, the hot–cold empathy gap (Chapter 7) renders such anticipations unlikely. Also, when they do occur, the effect may be the opposite one: "Let's not act now, while we are angry!" Plato is said to have asked another person to punish a slave who had misbehaved. According to Seneca, "His reason for not striking was the very reason that would have caused another to strike. 'I am angry,' said he, 'I should do more than I ought to and with too much satisfaction; this slave should not be in power of a master who is not master of himself.'" After 1945, some agents of transitional justice in Holland urged slowness, to avoid irreversible death sentences.

[22] In note 1, p. 139 I asserted that Seneca was mistaken when he claimed that invariance of intention over time was a proof that the agent was not governed by emotion. The converse statement seems correct, however.

a small recent crime was punished more severely than a grave crime committed earlier.[23]

Even more tellingly, the sentences that were meted out fitted the crimes very closely, in the sense that they reflected the action tendencies of emotions that were triggered by different crimes. The acts of informants and torturers induced *hatred*. According to Aristotle, the action tendency of this emotion is not (as in anger) to make the wrongdoer suffer, but to make him or her disappear from the face of the earth. These persons were in fact heavily represented among those who received the death penalty. Everyday acts of collaboration, such as joining the Nazi Party to keep one's job, were more likely to trigger *anger*. The legal expression of this emotion would be an ordinary prison sentence. When the wrongdoing was directed against a third party, as when Americans reacted to German atrocities against Soviet citizens, the emotion was one of (Cartesian) *indignation* rather than anger. The tendency for indignation to trigger weaker retribution than anger is reflected in the patterns of prosecution and sentencing. In the trials of German officers organized by the British and American occupational forces, high-level German war criminals, connected to mass murder, but who had not been involved in atrocities against British or American personnel, escaped prosecution, while the lowliest perpetrators who participated in beating, mistreating, or executing even a single prisoner of war were relentlessly pursued.

Finally, individuals whose collaboration took the form of diffusing pro-Nazi propaganda or being passive members of Nazi organizations were the objects of *contempt*. The action tendency of this emotion, ostracism, closely matched the legal reaction of imposing the loss of civil and political liberties, that is, a form of civic death. I find it hard to believe that these tight links between emotion-triggered action tendencies and forms of punishments could be accidental.

After the end of the war, many of those who remained neutral were blamed for their passivity, and were at the receiving end of angry or contemptuous reactions. Even if they were not harassed in any way, the guilt they felt for having done nothing may have strengthened their demand for retribution, as if post-transition aggression toward the wrongdoers could magically undo pre-transition passivity. This tendency for the neutrals, those in the "gray zone" between collaboration and resistance, to be especially vindictive, is an instance of transmutation (Chapter 9).

### Emotion and belief

Not only are most emotions triggered by beliefs; they can also affect belief formation directly as well as indirectly. The direct effect produces biased

[23] This fact might seem to go against the theory of the peak-end heuristic (Chapter 6).

beliefs, the indirect effect low-quality beliefs. One form of bias is illustrated in Stendhal's theory of *crystallization*. The origin of the term is as follows: "At the salt mines of Hallein near Salzburg the miners throw a leafless wintry bough into one of the abandoned workings. Two or three months later, through the effect of the waters saturated with salt which soak the bough and then let it dry as they recede, the miners find it covered with a shining deposit of crystal. The tiniest twigs no bigger than a tom-tit's claw are encrusted with an infinity of crystals, scintillating and dazzling." The analogy with love is clear: "From the moment you begin to be really interested in a woman, you no longer see her *as she really is*, but as it suits you to see her. You're comparing the flattering illusions created by this nascent interest with the pretty diamonds which hide this leafless branch of hornbeam – and which are only perceived, mark you, by the eyes of this young man falling in love."

In the saying by La Fontaine that I cited earlier and shall cite again, we easily *believe what we fear*. This, too, is a form of bias. In addition to the fact that we naturally (even in non-emotional states) may give excessive importance to low-probability risks, feelings of visceral fear may also cause us to believe that dangers are greater than they actually are. When we walk in a forest at night, a sound or a movement may trigger fear, which then causes us to interpret as fearsome other sounds or movements that we had previously ignored. The fear "feeds on itself."

The urgency of emotion acts on the gathering of information prior to belief formation rather than on the belief itself. The result is a low-quality belief, based on a less than optimal amount of information, but not a belief that is biased for or against any particular conclusion that the agent would like to be true. In practice, though, the two mechanisms often go together and reinforce each other. The agent initially forms an emotion-induced bias, and the urgency of emotion then prevents her from gathering the information that might have corrected the bias. As we saw in the previous chapter, wishful thinking is to some extent subject to reality constraints. Hence if the agent had gathered more information, it might have been difficult to persist in the biased belief.

A further cognitive effect of the emotions is that, while we are in their grip, it may be difficult to realize that they will subside (the "hot–cold empathy gap.") The shame-induced suicides I referred to earlier might not have occurred if the individuals had been able to anticipate that the contempt of others, and their own shame, would subside. People may also suffer from a "cold–hot empathy gap," which is the difficulty in anticipating, when in a calm state, the pains of a future experience such as being caught cheating on an exam or giving birth without anesthesia. The same persons might not have engaged in the behaviors that, when exposed, triggered the contempt of observers had they anticipated how horribly bad the shame would feel.

### Culture and emotions

Are all emotions universal? If not, are there some universal emotions? I answer a firm yes to the second question, and a tentative yes to the first.

It seems clear that some emotions are universal. It is commonly claimed that there exist half a dozen emotions – happiness, surprise, fear, sadness, disgust, and anger – that have facial expressions people recognize across cultures. Although the claim has been challenged, historians and anthropologists offer persuasive behavioral evidence. If one believes, as I do, that social norms exist in all societies, the emotions that sustain them – contempt and shame – must also be universal. One might imagine a society in which people felt anger when offended, but no (Cartesian) indignation when they observed offenses toward a third party. I find it hard to believe that such a society could exist, but I may be wrong. If love is universal, would not jealousy be too?

It is said that the Japanese have an emotion, *amae* (roughly rendered as helplessness and a desire to be loved), which does not exist in other societies. It has also been argued that ancient Greece was a "shame culture" that differed from modern "guilt cultures," that romantic love is a modern invention, and that the feeling of boredom (if that is an emotion) is of recent origin. One cannot exclude, however, that the allegedly absent emotions may have existed but not been conceptualized by members of the society in question. An emotion may be recognized as such by an external observer, but not acknowledged by the members of that society. In Tahiti, a man whose woman friend has left him will show the behavioral symptoms of sadness but will state only that he is "tired." In the West, the *concept* of romantic love is a relatively recent one, dating from the age of the troubadours. Prior to that time, there was only "merry sensuality or madness." Yet it is possible, and in my opinion likely, that the *experience* of romantic love occurred even when the society did not have the concept of that emotion. Individuals can be in love without noticing it, and at the same time their emotion may be obvious to observers, whether from their own society or from another. The ancient Greeks displayed a cluster of guilt-related reactions – anger, forgiveness, and reparations – that point to the presence of the emotion even if they did not have a word for it. The way people think about emotions may be culture specific, even if the emotions themselves are not.

One should add, though, that when a certain emotion is not explicitly conceptualized, it may also have fewer behavioral manifestations. La Rochefoucauld wrote that "some people would never have fallen in love if they had never heard of love." Guilt, too, may be more common in societies where people are told from an early age that they ought to feel guilty on this or that occasion.

### Bibliographical note

The best book on emotion is N. Frijda, *The Emotions* (Cambridge University Press, 1986). Aristotle's *Rhetoric* also provides invaluable insights into the causes and consequences of emotions. Seneca's *On Anger* goes beyond Aristotle in ways that prefigure the French moralists. I draw heavily on these writings in *Alchemies of the Mind* (subtitled *Rationality and the Emotions*) (Cambridge University Press, 1999), where the reader can find further references to the ideas discussed here. The idea of urgency, as distinct from impatience, is an addition to the framework of that book. I discuss it in more detail in "Urgency," *Inquiry* 52 (2009), 399–411. The book on *How Doctors Think* is by J. Groopman (New York: Houghton Mifflin, 2007). The references to McGeorge Bundy are from H. MacMaster, *Dereliction of Duty* (New York: Harper, 1997), p. 215, A. Preston, *The War Council: McGeorge Bundy, the ASC, and Vietnam* (Cambridge MA: Harvard University Press, 2006), p. 167, and D. Milne, *America's Rasputin: Walt Rostow and the Vietnam War* (New York: Hill and Wang, 2008), p. 149. For the role of surprise in emotional life, see B. Mellers, "Decision affect theory: emotional reactions to the outcomes of risky options," *Psychological Science* 6 (1997), 423–9. The experiment on escalation is in S. Shergill *et al*., "Two eyes for an eye: the neuroscience of force escalation," *Science* 301 (2003), 187. The loss-aversion theory of "two eyes for an eye" is transposed from D. Kahneman and A. Tversky, "Conflict resolution: a cognitive perspective," in K. Arrow *et al*. (eds.), *Barriers to Conflict Resolution* (New York: Norton, 1995), pp. 44–60. I discuss the issue in "Two for one? Reciprocity in Seneca and Adam Smith," *The Adam Smith Review* 6 (2011), 153–71. Sustained discussions of the role of emotions in explaining political behavior include P. Walcott, *Envy and the Greeks* (London: Aris and Phillips, 1978), R. Petersen, *Understanding Ethnic Violence: Fear, Hatred, and Resentment in Twentieth-Century Eastern Europe* (Cambridge University Press, 2002), and Chapter 8 of my *Closing the Books: Transitional Justice in Historical Perspective* (Cambridge University Press, 2004). I consider the Great Fear of 1789 and its effects on the constituent assembly in "The two great fears of 1789," *Social Science Information* 50 (2011), 317–29, and in "The night of August 4, 1789: a study of social interaction in collective decision-making," *Revue Européenne des Sciences Sociales* 45 (2007), 71–94. The ideas of "hot–cold" and "cold–hot" empathy gaps are due to George Loewenstein: see, for instance, "Emotions in economic theory and economic behavior," *American Economic Review: Papers and Proceedings* 90 (2000), 426–32. The claim that there are universal facial expressions of emotion is made by P. Ekman, "Facial expression and emotion," *American Psychologist* 48 (1993), 384–52, and criticized by L. Barrett, "Are emotions natural kinds?" *Perspectives on Psychological Science* 1 (2006), 28–58.

# 9 Transmutations

We are usually conscious of our desires and emotions. There are exceptions, to which I have referred and to which I shall refer again, but by and large these motivations operate in full transparency and with full awareness. By contrast, the motivations I shall refer to as (mental) *needs* operate "behind the back" of the agent.[1] He is unaware of them, and might deplore or resist their operation if he were aware of them. In the present chapter I catalogue and illustrate some of these needs and their effects. The term "transmutation" refers to the alchemy-like character of some of the operations I shall describe, not only the turning of lead into gold ("sweet lemons"), but also of gold into lead ("sour grapes"). Some operations also involve the creation of new mental states out of whole cloth. Although I have discussed some of these processes in earlier chapters and shall discuss others in later ones, I consider them together here to present a (relatively) full picture.[2]

The needs have causal efficacy because, when unsatisfied, they induce some kind of psychic discomfort that can be alleviated only by a mental alchemy. To verify this operation, which is unobservable in itself, we must consider the implications for observable facts, along the lines of the explanation of standing ovations on Broadway that appealed to the need for the reduction of cognitive dissonance (Chapter 1).

Let me first list the needs to be discussed, with in some cases the names of their foremost expositors:

> the need to justify one's choices by good reasons (Otto Neurath);
> the need for cognitive consonance (Leon Festinger);
> the need to believe that the world has meaning and order;
> the need for autonomy (Jack Brehm);

[1] This phrase, coined by Hegel, referred to the unintended consequences of social action (Chapter 17), but is equally apt in cases where the mind plays tricks on itself. Intentional action can be subverted by subintentional as well as by supraintentional causality.

[2] I shall use "motivation" as a general term to denote conscious desires and emotions, as well as unconscious needs. Hence when I refer to "motivated motivations," it is mostly in the sense of needs-generated conscious desires or emotions.

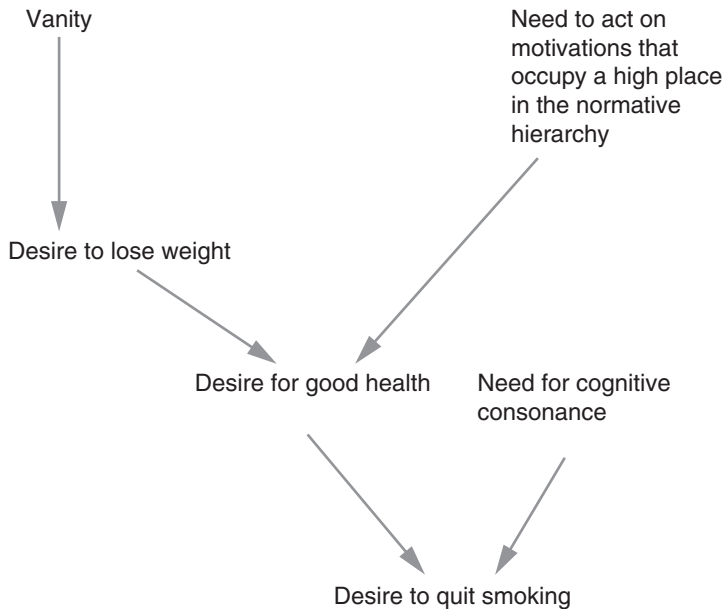



Figure 9.1

the need for novelty;
the need to maintain one's amour-propre (La Rochefoucauld);
the need to see oneself as guided by a motivation that is highly ranked in the hierarchy of motivations (Proust).

To introduce these ideas, I offer a concocted example. Suppose I am very vain, but fool myself into thinking that I am not. La Bruyère said that "Men are very vain, and hate to be seen as such." He probably meant seen *by others*, but one may reasonably extend his idea to an aversion to be seen as vain by oneself. Suppose also that my vanity makes me want to lose weight, but that my need to see myself as guided by a highly ranked motivation makes me believe I want to do so for reasons of health. However, because of my need for consonance, I cannot see myself as motivated by those reasons without also taking additional steps, such as to stop smoking. (If I acknowledged my vanity I would have no reason to stop and in fact have a reason to persist, since smoking makes it easier to lose weight.) The sequence can be shown in Figure 9.1.

As I said, the example is concocted, but it illustrates the principle of transmutation: under the normative pressure from the hierarchy of motivations, my vanity is transmuted into a concern for my health. It also illustrates the idea that one cannot fool oneself in an opportunistic manner. Doing so would entail

a loss of internal credibility, just as the opportunistic use in public of principled arguments – using them only when they coincide with the agent's interest – entails a loss of external credibility.

### The need to act for good reasons

As I note in Chapter 14, human beings desire to be rational. They also feel a need to have a *sufficient reason* for what they are doing. The desire and the need do not induce the same actions, since sometimes it *is rational to decide without a reason*, for instance by flipping a coin. Buridan's ass needed to have a reason for choosing one of two identical haystacks, and starved to death because he could not find one. Less dramatically, and more plausibly, Thomas Schelling tells of an occasion in which he had decided to buy an encyclopedia for his children. At the bookstore, he was presented with two attractive encyclopedias and, finding it difficult to choose between the two, ended up buying neither, despite the fact that had only one encyclopedia been available he would have bought it.[3] I give other examples in Chapter 14, where I also discuss how the need for reasons can give rise to the phenomenon of *hyperrationality*. Otto Neurath used the term "pseudorationalism" for the same phenomenon. Stating an idea to which I return in that chapter, he asserted that "[the] pseudorationalists do true rationalism a disservice if they pretend to have adequate insight exactly where strict rationalism excludes it on purely logical grounds. Rationalism sees its chief triumph in the clear recognition of the limits of actual insight."

### The need to reduce cognitive dissonance

I have already referred to the theory of cognitive dissonance, without going into details of its modus operandi. Although the word "cognitive" might suggest a conflict between two beliefs, the idea is more general. The coexistence of any two conscious states – beliefs, desires, emotions – can generate an unpleasant tension that can be resolved by adding or subtracting states or by changing one of the dissonance-creating states.[4] Leon Festinger, the originator of the theory, recounts that it was inspired by a conflict between a belief and an emotion. In the wake of an earthquake in India in 1934, the sociologist Prasad observed the puzzling fact of rumors of impending new disasters.[5] Festinger

[3] If he had chosen one at random, he might have come to see it as preferable, by attaching more importance to the dimensions – number of illustrations and of subjects covered, etc. – on which it was superior.

[4] The nature of this tension is rarely specified, and may vary from case to case.

[5] Since many of these rumors were about new earthquakes, they might arise from a rational understanding of the prevalence of aftershocks. Some of them, however, concerned cyclones, hurricanes and other phenomena with no rational connection to earthquakes. In his discussion of

stated the puzzle and its solution as follows: "Certainly the belief that horrible disasters were about to occur is not a very pleasant belief, and we may ask why rumors that were 'anxiety provoking' arose and were so widely accepted. Finally a possible answer to this question occurred to us – an answer that held promise of having rather general application: perhaps these rumors predicting even worse disasters to come were not 'anxiety provoking' at all but were rather 'anxiety justifying.'" In other words, the dissonance between the felt anxiety and the lack of obvious reasons for being anxious was reduced by the belief that there was in fact something to be anxious about.

This example is not compelling, since the increase in anxiety caused by the belief that something bad was about to happen would presumably more than offset the decrease in anxiety caused by the dissonance reduction. Be that as it may, the theory has offered many convincing explanations of puzzling phenomena.[6]

An example will show that the need for consonance can be closely related to the need to have reasons for acting. In a classical experiment, two groups of subjects are asked to write an essay offering arguments for the position on the pro-life versus pro-choice issue that they do *not* favor. The subjects in one group are paid a considerable sum of money for participating, whereas the others are asked to do so as a favor to the experimenter. After writing the essay, those in the second group but not those in the first display a more favorable attitude toward the position they have been arguing for. The explanation, plausibly, is that all the subjects desire to have a *reason* for what they are doing. Members of the first group can simply cite the money as their reason.[7] Members of the second group can cite their (adjusted) beliefs as the reason why they argue the way they do.[8]

the rumors of new disasters created by the great earthquake on July 21, 365 AD, Gibbon pointed to a different connection: the disaster revealed the anger of the gods, who could be expected to call down further calamities on the people. This mechanism may also have been at work in India.

[6] By and large, dissonance reduction is supposed to be an unconscious process. In the first statement of the theory, we find, however, a passage which imputes to the unconscious the capacity for deploying the kind of indirect strategy ("one step backward, two steps forward") that is the hallmark of the conscious mind: "What may one say concerning the seeking out of new information on the part of a person whose dissonance is near to the limit which can exist? Under such circumstances a person may *actively seek out, and expose himself to, dissonance-increasing information*. If he can increase the dissonance to the point where it is greater than the resistance to change of one or another cluster of cognitions, he will then change the cognitive elements involved, thus markedly reducing and perhaps eliminating the dissonance which is now so great" (my italics). He does not cite any examples, nor do I think there are any.

[7] Punishment for *not* participating would also be a sufficient reason. This explains why citizens under Communism might consistently have a system of double bookkeeping without their inner rejections of the system being undermined by their overt enthusiasm.

[8] In Pascal's wager, the reason for acting as if one believed is so overwhelmingly strong – the prospect of eternal bliss – that the believer need not look for another explanation of his behavior.

I have already mentioned the dissonance-reducing mechanism "It's so expensive (or unpleasant) that it must be good," illustrated by the standing ovations on Broadway, painful initiation rites in college, and (Proust's example) the visitors who persuade themselves that the sight of the sea and the sound of its waves are really enjoyable by taking an expensive hotel room with a view on the sea. I have also mentioned the tendency to upgrade – either before the choice or after – an option that is only slightly preferred to the others, to eliminate the unpleasant thought that a unchosen option might actually be superior. Rather than adding more examples from the scholarly literature on dissonance reduction, I shall discuss one *fictitious* case and one *fictional* case of tension between a desire to get X and a belief that X is unavailable. Such cases are ubiquitous, and the tensions to which they give rise can be very unpleasant. Often, the tension is resolved when the desire simply goes away. The agent may still *wish* he could be or have X, but he no longer has a causally efficacious want to get X. When some fifty years ago I realized that I did not have what it takes to be a first-class mathematician, my want to achieve that goal was simply extinguished. Yet I still think of it as perhaps the highest calling, as shown by the fact that I spend an absurd proportion of my time reading about mathematical achievements I only dimly understand.

Here I explore two reactions that take the form of alchemical transmutations. I shall do so by a fictitious example of a process as banal as it is common: rejected declarations of love or of marriage. I consider two persons, Peter and Ann. He wants to marry her; she refuses. We might then consider the scenarios in Figure 9.2.

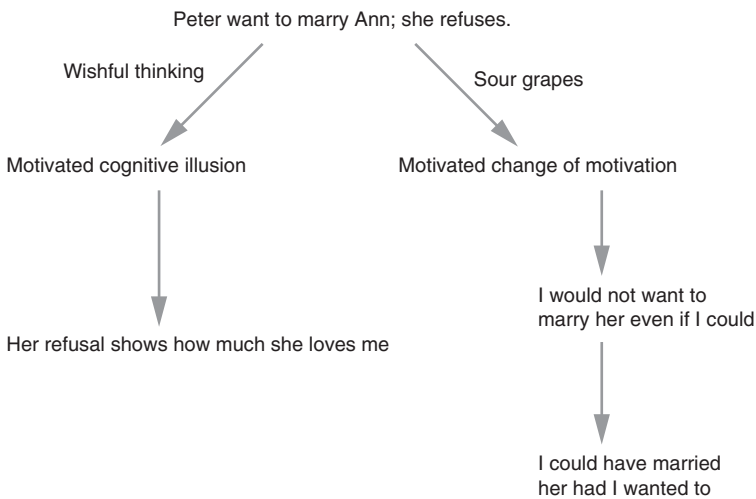



Figure 9.2

Peter's interpretation of Ann's refusal as a sign of her love for him is similar to how Stendhal, in his unrequited love for Méthilde Dembowski, interpreted her refusals. As his biographer writes, "So much anger, scruples, defense, severity [on her part] could only be proof of [her] love [for him]: one does not protest so strongly against anything but an all-powerful feeling."[9] This piece of wishful thinking, or perhaps self-deception, is manifestly irrational.

Yet this is not the only way of reducing dissonance. When there is a tension between a desire (or a judgment of desirability) and a belief, *something has to give* for the dissonance to be reduced, but what gives could be the desire as well as the belief. Which reaction is observed will presumably depend, among other things, on the ease of transmutation. In the example of the standing ovations on Broadway, the hard constraints on wishful thinking channeled the dissonance reduction into a different form, that of making the show appear as more enjoyable. In Peter's case, the constraints on wishful thinking might also be so hard that the more feasible reaction is a sour-grapes mechanism: downgrading her desirability as a marriage partner.

In itself, the downgrading is not irrational. As I will argue in Chapter 13, desires and preferences are not subject to assessments in terms of rationality or irrationality. Yet it might seem puzzling to claim that one transmutation, wishful thinking, is irrational, while another, sour grapes, is not. Since they are both induced by causal mechanisms operating "behind the back" of the agent, would not both be irrational? The answer depends on how we define rationality. One might indeed define it so that the sour grapes mechanism is no less irrational than wishful thinking, by emphasizing the *heteronomous* character of both transmutations. I believe, however, that for the explanatory purposes to which I shall harness the idea of rationality, this suggestion is unhelpful. Whereas the assumption of rationality can yield sharp predictions about what, in a given case, counts as a rational belief, the assumption of autonomy cannot.

Even if a sour-grapes reaction is not itself irrational, I believe it is often accompanied by irrational belief formation. Having taken the first step of making Ann appear to be an undesirable marriage partner, Peter might need to free himself from the obvious self-suspicion of being subject to a sour-grapes reaction by taking the further step of persuading himself that he could easily have married her had he wanted to.[10] As an independent example of this

[9] More recently, Freudian psychology has contributed to the tendency to refuse to take answers to questions or accusations at face value. Someone who reacts angrily to an unfounded accusation may see his anger interpreted as evidence for the charge: every protest is "too much." This tendency may perhaps count among the *harms* caused by the theory (see the Conclusion).

[10] According to Aristotle, Thales of Miletia sought to liberate himself from the same suspicion by the more robust means of *taking action*: "[Thales] was reproached for his poverty, which was supposed to show that philosophy was of no use. According to the story, he knew by his skill in

pattern – that is, not a fictitious one invented for the purpose of making a conceptual point – I shall offer a fictional case, Proust's analysis of the richly comical character Legrandin, whose outwardly anti-snob attitude hides deep inward snobbery.

The Narrator cites his grandmother's surprise at "the furious invective which [Legrandin] was always launching at the aristocracy, at fashionable life, and 'snobbishness' – 'undoubtedly'. He would say, 'the sin of which Saint Paul is thinking when he speaks of the sin for which there is no forgiveness.'" From the context, it seems that the grandmother thought that he "doth protest too much." If so, this impression is confirmed later, when the Narrator innocently asks Legrandin whether he knows the Guermantes family, the epitome of the high aristocracy. The acuity of the Narrator's analysis of Legrandin's response justifies, I hope, a lengthy quotation:

> [At] the sound of the word Guermantes, I saw in the middle of each of our friend's blue eyes a little brown dimple appear, as though they had been stabbed by some invisible pinpoint, while the rest of his pupils, reacted by secreting an azure overflow. His fringed eyelids darkened, and drooped. His mouth, which had been stiffened and seared with bitter lines, was the first to recover, and smiled, while his eyes still seemed full of pain, like the eyes of a good-looking martyr whose body bristles with arrows. "No, I do not know them," he said, but instead of uttering so simple a piece of information, a reply in which there was so little that could astonish me, in the natural and conversational tone which would have befitted it, he recited it with a separate stress upon each word, leaning forward, bowing his head, with at once the vehemence which a man gives, so as to be believed, to a highly improbable statement (as though the fact that he did not know the Guermantes could be due only to some strange accident of fortune) and with the emphasis of a man who, finding himself unable to keep silence about what is to him a painful situation, chooses to proclaim it aloud, so as to convince his hearers that the confession he is making is one that causes him no embarrassment, but is easy, agreeable, spontaneous, that the situation in question, in this case the absence of relations with the Guermantes family, might very well have been not forced upon, but actually designed by Legrandin himself, might arise from some family tradition, some moral principle or mystical vow which expressly forbade his seeking their society.
>
> "No," he resumed, explaining by his words the tone in which they were uttered. "No, I do not know them; I have never wished to know them; I have always made a point of preserving complete independence; at heart, as you know, I am a bit of a Jacobin.

the stars while it was yet winter that there would be a great harvest of olives in the coming year; so, having a little money, he gave deposits for the use of all the olive-presses in Chios and Miletus, which he hired at a low price because no one bid against him. When the harvest time came, and many were wanted all at once and of a sudden, he let them out at any rate which he pleased, and made a quantity of money. Thus he showed the world that philosophers can easily be rich if they like, but that their ambition is of another sort." In Montaigne's version, when Thales condemned money making, he "was accused of sour grapes like the fox."

People are always coming to me about it, telling me I am mistaken in not going to Guermantes, that I make myself seem ill-bred, uncivilized, an old bear. But that's not the sort of reputation that can frighten me; it's too true!"

If I asked him, "Do you know the Guermantes family?" Legrandin the talker would reply, "No, I have never cared to know them." But unfortunately the talker was now subordinated to another Legrandin, whom he kept carefully hidden in his breast, whom he would never consciously exhibit, because this other could tell stories about our own Legrandin and about his snobbishness which would have ruined his reputation for ever; and this other Legrandin had replied to me already in that wounded look, that stiffened smile, the undue gravity of his tone in uttering those few words, in the thousand arrows by which our own Legrandin had instantaneously been stabbed and sickened, like a Saint Sebastian of snobbery: "Oh, how you hurt me! No, I do not know the Guermantes family. Do not remind me of the great sorrow of my life." And since this other, this irrepressible, dominant, despotic Legrandin, if he lacked our Legrandin's charming vocabulary, showed an infinitely greater promptness in expressing himself, by means of what are called "reflexes," it followed that, when Legrandin the talker attempted to silence him, he would already have spoken, and it would be useless for our friend to deplore the bad impression which the revelations of his alter ego must have caused, since he could do no more now than endeavor to mitigate them.

*À la recherche du temps perdu* offers many other examples of the transmutation of "I cannot have it" to "I do not want to have it."[11] Although it does not seem that Proust was aware of Nietzsche as a predecessor,[12] his description is strikingly similar to the latter's idea of a "transvaluation of all values." Nietzsche describes the "workshop" of this alchemy as follows:

It is a careful, crafty, light rumor-mongering and whispering from every nook and cranny. It seems to me that people are lying; a sugary mildness clings to every sound. Weakness is going to be falsified into something of merit . . . And powerlessness which does not retaliate is being falsified into "goodness," anxious baseness into "humility," submission before those one hates to "obedience" (of course, obedience to the one who, they say, commands this submission – they call him God). The inoffensiveness of the weak man – cowardice itself, in which he is rich, his standing at the door, his inevitable need to wait around – here acquires a good name, like "patience," and is called virtue itself. That incapacity for revenge is called the lack of desire for revenge, perhaps *even forgiveness.*

[11] The opposite effect can be observed in schoolchildren when they transmute "I do not want to do my homework" into "My homework is so difficult that I cannot do it." In such cases, Seneca observed, "The reason is unwillingness, the excuse, inability."

[12] Both Nietzsche and Proust, however, were influenced by La Rochefoucauld. They may well have read the following Maxim: "The scorn for riches displayed by the philosophers was a secret desire to recompense their own merit for the injustice of Fortune by scorning those very benefits she had denied them; it was a private way of remaining unsullied by poverty; a devious path toward the high respect they could not command by wealth." Note, however, that this explanation of the philosopher's scorn for riches does not apply to Thales.

### The need to find meaning and order in the world

People have a strong need to find meaning, order, and patterns in the world. As the giveaway phrase goes, "It is no accident that …" The mind abhors accidents. I shall focus on three variations on this theme: the imputation of intentional agency, the use of objective teleology, and the reliance on analogy.

The tendency to seek an explanation of events in terms of an *intention* to bring them about is pervasive, even when – as in the cases I shall discuss – there is no *evidence* for intentional agency. Conspiracy theories offer numerous examples. Three caveats are necessary. First, history shows many examples of actual conspiracies, such as the Gunpowder Plot or the Babington Plot (the conspiracy to assassinate Queen Elizabeth I). Hence some conspiracy theories may be true, and others, if not true, may be held rationally. Second, some conspiracy theories are entertained for their consumption value or entertainment value, without the full mental endorsement that is needed if they are to serve as premises for action. Third, the propagators of such theories may know that they are false, yet spread them in the expectation – which may be quite rational – that the recipients will believe them. Thus, assuming the leaders of Hamas to be minimally rational and well informed, they probably do not believe in Article 22 of their Charter, stating that Jews were "behind the French Revolution, the Communist revolution and most of the revolutions we heard and hear about, here and there. With their money they formed secret societies, such as Freemasons, Rotary Clubs, the Lions and others in different parts of the world for the purpose of sabotaging societies and achieving Zionist interests. With their money they were able to control imperialistic countries and instigate them to colonize many countries in order to enable them to exploit their resources and spread corruption there." It seems plausible, nevertheless, that the general public in the Middle East sincerely believes in the existence of a Zionist conspiracy and that many, for this reason, are willing to kill or be killed in fighting it.

Although the best-known study of conspiracy theories in politics is titled "The paranoid style in American politics," I shall cite an example from French history that I know better. In eighteenth- and early nineteenth-century France, the population could never accept that a dearth of grain might be due only to bad weather and a bad harvest. In the words of Georges Lefebvre, "The people were never willing to admit that the forces of nature alone might be responsible for their poverty and distress"; instead they sought an explanation in terms of agency. The usual assumption was that hoarders, who were indifferent to the welfare of the people, had bought grain to drive the price up. Sometimes, the lack of grain was even explained in terms of the desire of malevolent elites to starve the people, as part of ongoing class warfare. In the Great Fear that swept over France in July 1789, many interpreted the fact that it broke out

simultaneously and independently in six or seven regions as evidence of an aristocratic conspiracy. The inference to a common cause was correct, but the cause was misidentified. Since the dearth of grain was especially acute just before harvest time and since harvest took place more or less at the same time throughout France, this common cause produced the same effect everywhere. In a similar episode from the early nineteenth century, the government exhibited the same mindset, when it took the simultaneous eruption of similar rumors as evidence of a nationwide conspiracy to drum up popular unrest. Once again, there *was* a common cause – the nationwide harvest.

Commenting on an episode in the English Civil War, Hume observed that "whatever hardships [the troops] underwent, though perhaps derived from inevitable necessity, were ascribed to a settled design of oppressing them." This tendency to interpret causal necessity as intentional design is often observed in popular and even professional views of the causes of cancer and other diseases. Two steps may be involved. First, there is well-documented tendency (Chapter 7) to perceive patterns in random events. Second, there is the tendency to seek the cause of patterns, whether real or spurious, in *agency.* Rather than looking for non-human causes of cancer, such as the emission of radon from the earth, we first look for the cause in pollution, industrial waste, and pesticides. Commenting on this tendency, a writer on cancer refers to the "comfort in believing that humans, through their own devices, have increased the likelihood of cancer. What free-willed creatures have created can conceivably be undone. Failing that, there is at least a culprit to blame." The last sentence is consistent with my argument here. The preceding sentence suggests an explanation in terms of wishful thinking (Chapter 7) rather than the need for meaning. The writer adds that the explanations in terms of man-made causes "fit so perfectly with the rest of our world-view that there was little incentive to look deeper … If we or someone we knew ever got cancer we were quick to wonder whether corporate America was to blame."

*Natural and social scientists are people too*, with the same cognitive needs as the rest of us. In 2010, a White House advisory group issued a report on *Reducing Environmental Cancer Risk*, which concluded that the number of cancer cases associated with industrial carcinogens "has been grossly underestimated." Critics charged that the report ignored other major environmental causes of cancer – smoking, diet, infection, geophysical sources of radiation, and the like – to focus on pollutants and occupational exposures. The issues are technical and controversial, and well beyond my competence. It seems, though, that the report is heavily biased toward speculation about unknown but possible man-made causes.

In his history of the French *ancien régime*, Alexis de Tocqueville also fell into an "agency trap" when he imputed a divide-and-conquer strategy to the successive kings. He observed, correctly, that the tax exemption granted to the

nobles removed any occasions on which they might make common cause with the bourgeoisie, and asserted, also correctly, that this fragmentation of the elite worked out to the benefit of the kings. He also claimed that the "kings were able to create [their] unchecked power only by dividing the classes and isolating each of them amid its own peculiar prejudices, jealousies, and hatreds so as never to have to deal with more than one at a time." He did not even try, however, to substantiate the alleged agency of the kings. He seems to have inferred agency and intention from objective benefits. His need to find meaning and patterns made him overlook the role of *non-explanatory* or *accidental benefits*. Agency provides meaning; accidents do not.

By "objective teleology" I mean the tendency to explain events or facts in terms of their objective consequences (or alleged consequences). This tendency can be hard to distinguish from the previous one, because, as I just noted, we often infer subjective intentions from objective consequences, especially if the latter work out to the benefit of an agent.[13] Yet sometimes the mere observation of benefits seems to suffice for explanation. The fact that wars benefit the munitions industry has often been used, and not only in the novels by Eric Ambler and Graham Greene, to support the idea that the activities of the "merchants of death" *explain* wars. Maybe they do, but the benefits in themselves prove nothing. Another instance of the tendency is the idea of "objective complicity," a form of strict liability. The best-known examples are from the Stalinist period, when individuals were shot for voicing opinions that objectively (in the subjective view of the leader) served the interests of the class enemy. After the "événements" in May 1968, Jean-Paul Sartre used the same argument *against* the Stalinists, when he accused the French Communist Party of objective complicity with de Gaulle's government. Speaking after the Watts riots in 1965, Martin Luther King asserted that "the purpose of the slum is to confuse those who have no power and perpetuate their powerlessness." Although he may simply have intended to identify the *effects* of the slum, the teleological language is symptomatic.

This tendency to explain by consequences may itself have a deep-seated explanation. According to one summary, "a ready recourse to [teleological] explanations has long been observed in young children and has been thought to be increasingly restrained in us as our acquisition of causal mechanical ideas and knowledge proceeds. However, recent research suggests that the

[13] The inference can also be made when the alleged consequences *harm* someone else. Thus in 1962, William Buckley warned the leader of the John Birch Society against accusing American leftists of wanting to bring about a Communist America: " I hope the Society thrives, provided, of course, it resists such false assumptions as that a man's subjective motives can automatically be deduced from the objective consequences of his acts." In his view, some leftists were misguided, not ill-intentioned.

inclination to these explanations is never truly overcome but instead remains a default mode among adults and becomes salient again in patients with Alzheimer's disease." For instance, under pressure to give a quick response, college-educated adults accepted the claim that "fungi grow in forests to help decomposition." They will not go as far as to accept that we have noses to provide support for our glasses or two kidneys to allow one of them to be transplanted, but less obviously absurd claims may provide a "mental click" that is easily confused with the click of explanation.

In this respect, too, (some) social scientists behave like the people who (other) social scientists are studying. The objective teleology then takes the form of *functional explanation*, a mode of scientific reasoning I often criticize in this book. In *The Theory of Moral Sentiments*, Adam Smith frequently appeals to objective teleology, as when he says that "every thing is contrived for advancing the two great purposes of nature, the support of the individual, and the propagation of the species." In his vaguely deistic mode of thinking, it is possible that the verb "contrived" had a divine subject. In other cases, the explanation is offered in an unambiguous agent-less mode, citing *free-floating intentions* and *verbs without subjects*. A text by Michel Foucault on the effects of the prison system is representative:

> But perhaps one should reverse the problem and ask oneself what is served by the failure of the prison: what is the use of these different phenomena that are continually being criticized; the maintenance of delinquency, the encouragement of recidivism, the transformation of the occasional offender into a habitual delinquent, the organization of a closed milieu of delinquency. Perhaps one should look for what is hidden beneath the apparent cynicism of the penal institution, which, after purging the convicts by means of their sentence, continues to follow them by a whole series of "brandings" (a surveillance that was once de jure and which is today de facto; the police record that has taken the place of the convict's passport) and which thus pursues as a "delinquent" someone who has acquitted himself of his punishment as an offender? Can we not see here a consequence rather than a contradiction? If so, one would be forced to suppose that the prison, and no doubt punishment in general, is not intended to eliminate offences, but rather to distinguish them, to distribute them, to use them; that it is not so much that they render docile those who are liable to transgress the law, but that they tend to assimilate the transgression of the laws in a general tactics of subjection. Penality would then appear to be a way of handling illegalities, of laying down the limits of tolerance, of giving free rein to some, of putting pressure on others, of excluding a particular section, of making another useful, of neutralizing certain individuals and of profiting from others.

If we translate the rhetorical questions and insinuating statements into ordinary assertions, the text illustrates how the need for meaning can transform prison failure, which may have occurred for all sorts of accidental reasons, such as lack of resources and bad planning, into a seamless pattern of *oppression without oppressors*. Teleology provides meaning; accidents do not.

Finally, the "hermeneutics of suspicion," to which I referred in Chapter 4 and shall discuss again in Chapter 16, illustrates objective teleology in the interpretation of texts. The benefits that accrue to a character in a novel serve to explain her behavior. The impression that a text makes on the reader serves to explain the strategy of the author.

*Analogies*, like agency and benefits, can also produce a mental click that is easily confused with the click of explanation. We notice that a situation or event X has property A, that another situation or event Y has property A *and* property B, and infer that object or event X also has property B. Because of what might be called the first law of pseudo-science – *everything is a little bit like everything else* – the mind will almost always be able to find some similarity. The property B may be stated either as a factual or as a normative proposition. In the latter case, the fact that doing B was successful (or unsuccessful) in situation Y is used to argue directly for (or against) doing B in X. In the former case, the belief that X has factual property B may be used as a premise for action (or inaction). Often, however, the statement that X has factual property B is the end of the story – no action is called for.

Although I shall focus on how analogies can provide meaning and order in the social or physical universe, it is not always easy to separate this role from other psychological functions of analogies.

(i) In some cases, analogies are used instrumentally, to persuade others of a conclusion that the agent has reached on other grounds. I shall ignore this case, to focus on analogies that are causally efficacious in shaping or supporting the agent's own beliefs.

(ii) In other cases, pure wishful thinking (Chapter 7) is at work. We choose analogies that support the conclusion we would like to be true. The conclusion triggers the search for an analogy that supports it, by a characteristic form of backward reasoning.

(iii) In still other cases, we use an analogy to bolster a conclusion that we have reached on other grounds. If a "bottom-up" (evidence-based) analysis yields a conclusion as somewhat probable, but not certain, we may experience cognitive dissonance that can be reduced by supplementing the evidence with a "top-down" analogy. The examination of the evidence generates a conclusion, the uncertain or tentative character of which triggers the search for an analogy.

(iv) In the cases I consider here, an analogy spontaneously comes to mind, and we adopt it without examining the evidence, that is, without asking whether X really has property B. Once we have adopted it, we tend to ignore evidence that might suggest that X does not have property B.

The causes of spontaneous adoption of analogies vary. In the most thorough analysis of the analogies used to argue for and against various options during

the American war in Vietnam, the author cites the availability and representativeness heuristics (Chapter 14), the recency effect (Chapter 6), and "surface commonalities" or, as cognitive psychologists call them, "mere appearance." Because surface similarities provide order and meaning, they can trigger a spurious form of satisfaction. Commenting on alleged parallels between successive episodes in pre-revolutionary France, an historian observes that "the repetition of motifs and patterns ... may provide a *pleasing* interpretative tapestry, but it is basically unhistorical" (my italics). By implication, the proper historical approach is to disencumber the mind, as far as possible, of ready-made schemata. An analytical philosopher described the mindset of a non-analytical author as "essentially *magical* thought, the primitive conception that similarities must point to powers, and analogies in thought stand for a kind of causation" and added that "as magical, it is also at a deep level *comforting*" (my italics). He preferred the bleakness of analytical philosophy.

In any given case, several causes may be at work and reinforce each other. The availability and recency heuristics may suggest analogies whose appeal is enhanced by surface commonalities; conversely, lack of such commonalities may undermine them. I shall offer some examples from the discussions among decision makers during the American war in Vietnam that illustrate both possibilities. Although the appeal to analogies is common in many wartime situations, the Vietnam War seems to have generated an exceptionally large number of analogies (see Table 3.1).

Robert McNamara's analogy with the Cuban crisis, which represented what an historian calls "his only real experience with the planning and direction of military force," was not taken seriously in private discussions (see Table 3.1). Similarly, few (except for President Johnson) listened to Walt Rostow's argument for bombing the oil storage facilities in North Vietnam, based on an analogy with his own experience in planning the strategy of bombing German oil facilities at the end of World War II. Nor does anyone seem to have listened to the Chairman of the Joint Chiefs of Staff, Earl Wheeler, when he argued that the Tet Offensive was only a temporary setback, on an analogy with the Battle of the Bulge (where he had served). By contrast, the Korean analogy, which the availability heuristic triggered in many decision makers with experience from the Korean War, seemed to offer enough surface similarities to make a "pleasing tapestry." Intrinsically, as George Ball pointed out, the analogy was severely halting, yet the similarities caused other members of the Johnson administration to dismiss the differences as "nuances."

The law, too, relies heavily on analogical thinking. Montaigne observed that "Just as no event and no form completely resembles another, neither does any completely differ ... All things are connected by some similarity; yet every example limps and any correspondence which we draw from experience is always feeble and imperfect; we can nevertheless find some corner or other by

which to link our comparisons. And that is how laws serve us: they can be adapted to each one of our concerns by means of some twisted, forced or oblique interpretation." He seems to imply that the "concerns" (*affaires*) precede and motivate the analogies, as in case (ii) above. This pattern certainly occurs, but there is no reason to think that legal analogies are always motivated. Consider the decision by the US Supreme Court in *Tanner* v. *United States* (1987). The Court considered a jury trial in which seven jurors had regularly consumed alcohol during the noon recess, and found that this fact did not constitute grounds for overturning the verdict: "However severe their effect and improper their use, drugs or alcohol voluntarily ingested by a juror seems no more an 'outside influence' than a virus, poorly prepared food, or a lack of sleep." I have no reason to think – although it is possible – that the choice of these far-fetched analogies was dictated by the desire to reach a particular conclusion. They may just have occurred spontaneously to the judges. The Court could equally well, however, have focused on a *difference* between ingesting alcohol and catching a virus, namely that only the first is done voluntarily. As suggested by the quotation from Montaigne, *the second law* of pseudo-science could be, "Everything is a little bit different from everything else."

Analogies may satisfy a need. Intellectually, they are worthless, unless the hypotheses they inspire are examined on their merits. They can cause great harm when used as premises for policy or legal decisions. Politicians can certainly learn from history, but not by identifying macro-analogies. Instead, they should study history to understand human nature and the particularities of local conditions. If American politicians had known that China was the historical archenemy of Vietnam, they would not have used the illusion of a monolithic Communist bloc as a premise for their actions in 1963–1964 and later.[14] Rather than being guided by the first law of pseudo-science, analysts and decision-makers should adopt Bishop Butler's dictum, "Everything is what it is and not another thing," or follow William Blake's maxim, "Art and Science cannot exist but in minutely organized particulars." Yet particulars do not offer meaning and coherence; analogies do.[15]

[14] The Asia specialists in the State Department told them so, but were overruled by CIA and the Department of Defense, as well as by the head of their department. In the words of one historian, "Talk of hereditary enemies was nothing [Secretary of State] Rusk cared to hear, for it did not fit the cold war scenario, in which such old-fashioned items as nationalism and geography faded into the deep background." The Americans made the same mistake about the Vietnamese as the English had made about the Americans two centuries earlier. In the words of David Ramsay, writing in 1789, "the British supposing the Americans, to be influenced by the considerations which bias men in the languid scenes of tranquil life, and not reflecting on the sacrifices which enthusiastic patriotism is willing to make," thought they could bring the colonies to their knees by devastating their possessions.

[15] Also, the path of particularism is thorny and painful. Absorbing the history, culture and language of another country takes a decade or more. As Tocqueville noted, "general ideas . . .

Once again, scientists, too, can be affected by this form of pattern seeking. The cabinet of horrors of science is replete with analogies.[16] The analogy with society and biological organisms, for instance, has been used to support the idea that societies, like organisms, are self-regulating entities with built-in homeostatic correction mechanisms, such as revolutions. In the nineteenth century, scholars debated what, in society, would correspond to the cell in the organism, without asking themselves whether there was any reason to expect any analogy at all. The explanation of market competition by an analogy with natural selection has failed dismally (Chapter 11). A currently fashionable analogy is between genes and "memes." Other writers have used physical rather than organic analogies and looked for the social equivalent of Newton's laws or the force of gravity. Bertrand Russell said that "the fundamental concept in social science is Power, *in the same sense* in which Energy is the fundamental concept in physics" (my italics). Physicists such as Niels Bohr and Ilya Prigogine tried, unsuccessfully, to apply their ideas to social phenomena. Scholars who argue that the social sciences can have an impact on the object they study routinely invoke Heisenberg's uncertainty principle, as if the profundity of his principle could turn their truism into something equally profound. Within the social sciences, too, scholars have tried, unsuccessfully, to graft ideas from one discipline onto another. Examples include analogies between power and money, between phonemes and "mythemes," between gothic cathedrals and scholastic treatises, between individual precommitment and constitution making, and between physical capital and social capital.[17]

I am not denying that analogies can be useful in suggesting hypotheses. The Rutherford–Bohr model of the atom on an analogy with the solar system was useful as a ladder, which later physicists could kick away behind themselves. The idea of viewing the genetic material on an analogy with language proved successful, although some hypotheses suggested by the analogy were falsified. The analogy between utility or profit maximization and the maximization of biological fitness led to a fruitful marriage of game theory and biology. However, the origin of a fruitful hypothesis is irrelevant for its validity. Ideas are to be judged by their descendants, not by their ancestors.

make it unnecessary to waste time delving into particular cases." This source of the attraction of analogies must be distinguished, however, from their ability to satisfy the need for meaning.

[16] Wikipedia has an entry on "parallelomania," which reads in part as follows: "In historical analysis, biblical criticism and comparative mythology, parallelomania refers to a phenomenon where authors perceive apparent similarities and construct parallels and analogies without historical basis. The concept was introduced to scholarly circles in 1961 by Rabbi Samuel Sandmel."

[17] These analogies have been offered by, respectively, Talcott Parsons, Claude Lévi-Strauss, Pierre Bourdieu, me, and Robert Putnam.

### The need for autonomy

Earlier, I referred to the phenomenon of *reactance*. Like its cousin, cognitive dissonance reduction, it relies on psychological mechanisms that can be hard to identify with precision. Here, I first cite a number of instances of reactance, beginning with one already mentioned, and then discuss the possible causes. Except for the last two examples, the cases are taken from the psychological literature on reactance, either from laboratory experiments or from field studies.

(1) When a choice option is eliminated, subjects tend to rank it more highly than before.[18]
(2) When toddlers can see comparable toys behind two barriers, one low and another higher, they show more interest in the toys behind the high barrier.[19]
(3) Similarly, children find candies more attractive when they are placed further away from them rather than within their reach.
(4) When seats belts are made mandatory, some drivers are less likely to use them.[20]
(5) When subjects in a swimming-pool environment are given sheets saying either "Don't litter" or "Help keep your pool clean," a larger proportion of those who received the first injunction littered the sheets rather than putting them in a trash can.
(6) When people make phone calls from a public phone booth, they tend to make longer calls if they can see that another person is waiting to make a call.
(7) Similarly, when drivers are about to leave a parking lot, they tend to do so more slowly if they can see that other drivers are circling to take their spot.
(8) In some New York taxis, customers who use credit cards are presented with a screen that provide them with the options of tipping 20 percent, 25 percent or 30 percent – well above normal tipping rates. Compared to

[18] The laboratory demonstration of this effect cited in Chapter 2 might be replicated in field studies. Thus when a presidential candidate retires from the race, his or her popularity score should go up.

[19] The effect was observed only for boys, not for girls, suggesting that, for whatever reason, these two-year-old girls were less motivated than boys by challenges.

[20] Other drivers are of course more likely to do so. They may, however, be subject to a different counterproductive effect, if the enhanced safety makes them drive more recklessly. For this reason, Chicago-style economists have objected to the mandatory use of seatbelts. Because of their tendency to see preferences as fixed and stable, they have not cited reactance as an objection.

"traditional" taxis, these are twice as likely to receive a tip of zero. At the same time, the average size of tips increases by about 10 percent.

(9) Some jurors refuse to ignore evidence that the judge finds inadmissible because he *tells* them to ignore it.

(10) Some patients refuse to take their medication because the doctor *tells* them to take it.

(11) Some doctors do not adhere to the recommendations of medical authorities. According to one study, doctors' "resistance to imposed activities (cookbook medicine)" is an important barrier to the implementation of diabetes guidelines in the Netherlands.

(12) The psychoanalyst Jacques Lacan refers to "that resistance of *amour-propre* [unlike the resistance of the unconscious], to use the term in all the depth given to it by La Rochefoucauld, and which is often expressed thus: I can't bear the thought of being freed by anyone other than myself."

(13) In a biography of François Mitterrand, we read that, "In the 1950s, as Interior Minister, he had once received the Algerian nationalist leader, Ferhat Abbas. After Abbas had waited in an anteroom for an hour and a half, an aide went in to find the cause of the delay. Mitterrand was reading the cartoons in *France Soir*. It was not that he had intended a political snub. Nor was it just rudeness or thoughtlessness or egoism – though it was certainly all those things too. The explanation was less rational. He had a visceral reaction against any kind of restriction – whether in politics, private life or the realm of ideas. Punctuality was a straitjacket he refused to accept."

Everyday language has a word for such reactions: *willful*, defined by the Oxford English Dictionary as "asserting or disposed to assert one's will against persuasion, instruction, or command." We have all observed this phenomenon, in others and in ourselves. The puzzle is *why* people form these preferences (cases 1–3) or choose these behaviors (cases 4–13). *What's in it for the person?* According to the psychological theory of reactance, when an individual's freedom to choose is eliminated or threatened, she is motivated to restore or confirm the freedom, resulting in enhanced attraction of eliminated alternatives and a tendency to choose threatened alternatives. Whereas all the examples fit this abstract account, the concrete mechanisms differ.

In the first three examples, no interpersonal interaction is involved. In case (1), the subjects in the "choice" condition were told that their third-ranked option had been eliminated "for some unknown reason," hence they had no reason to *resent* the restriction of the choice set. They might nevertheless *deplore* the loss of an option, even one they would not have chosen. As Isaiah Berlin wrote, "It is the actual doors that are open that determine the extent of

someone's freedom, and not his own preferences." If people value freedom in this sense, they might indeed react to the loss of even a low-ranked option. The puzzle is why the reaction would take the form of upgrading the value of that option. To my knowledge, the psychological literature does not address that question, and I cannot propose an answer.

In cases (2) and (3), the explanation seems to be (I am reconstructing it) that the scenarios implicitly suggest that the child will take the easiest path to the goal, and that he reacts to that suggestion as an obstacle to his autonomy. If this is indeed the correct explanation, it offers an interesting contrast with the principle of rational choice, which can be seen, at least in part, as a generalization of the principle of least effort (Chapter 13).

Cases (4)–(13), which do rely on interpersonal interactions, are easier to understand: *people do not like to feel crowded*, and are willing to give up some other good to avoid the feeling. The person in the phone booth or the driver leaving the parking spot might, if nobody was waiting for their place, have acted more expeditiously and achieved her ends (e.g. getting back in time for dinner) better. Yet knowing that others want them to leave as soon as possible, they slow down to convince themselves that they are not acting under pressure, but freely and autonomously. The other examples provide variations on this basic theme, the most interesting being perhaps that of Mitterrand, who felt his autonomy threatened by an appointment he had presumably made himself in a fully autonomous way. He behaved like the witty Count Chalvet in Stendhal's *Le rouge et le noir*: "'Oh, I am an Independent' he was saying to a gentleman wearing three badges whom he was evidently mocking. 'Why should I be of the same opinion today that I was six weeks ago? Why, my opinion would be my tyrant.'"

### The need for novelty

Desires can also be affected by an unconscious need for novelty or change. In H. C. Andersen's tale "What Father Does Is Always Right," the farmer goes to the market in the morning to sell or exchange his horse. First, he meets a man with a cow, which he likes so much that he exchanges the horse for it. In successive transactions, the cow is then exchanged for a sheep, the sheep for a goose, the goose for a hen, and, finally, the hen for a sack of rotten apples. The farmer's road to ruin is paved with stepwise improvements. Each time the farmer believes himself to be better off by the exchange, but the net result of all the exchanges is disastrous.

More formally, imagine a person who regularly (although not consciously) adjusts his desires so that he prefers more strongly the commodity of which he currently has less. Suppose he is exposed to the following sequence of two-commodity bundles: (1/2, 3/2), (3/4, 1/2), (1/4, 3/4), (3/8, 1/4) . . . Then if at a

given time he is consuming bundle n in the sequence and for the next period is offered the choice between bundle n and bundle n + 1, he will always choose the latter since it offers more of the commodity of which he currently has less. But since the sequence converges to zero, these local improvements bring about overall ruin. The effect is similar to that in which an agent can be bled to death by cycling preferences (Chapter 13), but the mechanism is different.

We can also observe the need for novelty in artists (and the publics of artists) who confuse creativity and originality. Rather than following the imperative "Enchant me!," artists respond to Diaghilev's command "Astonish me!" The use of randomization in the arts is but one of many examples of the sterile search for originality. The search may, to be sure, be an effect of one-upmanship rather than responding to an inner need, but I conjecture that in many cases it is a compensatory response to a failure of creativity.

### The need to maintain amour-propre

Here I shall add a few comments to the discussion of amour-propre, self-love, in Chapter 5. As I understand the idea of amour-propre discussed by the seventeenth-century French moralists, it refers to a person's need to maintain a good image of herself, in her own eyes and in the eyes of others. In other words, the person constantly acts for an external or an internal *audience*. Whereas the idea of an external audience is clear enough, that of an internal audience is more like a metaphor. It is related to the "warm glow" that I discussed previously, the pleasure we feel when we do something to benefit others, unbeknownst to them and in the absence of observers.

The search for the approval of an external audience is easy to document. In Chapter 5 I cited Bentham's dismissive reference to claims that the refusal of Necker, the minister of Louis XVI, to be paid for his services was "a sophisticated means to satisfy his greed." As all his contemporaries knew, and as even his adoring daughter Mme de Staël admitted, his show of disinterestedness was due only to his monumental vanity. George Washington, too, desired intensely to be viewed as disinterested, although when unobserved he departed from his high standards. One may have instrumental reasons to build a reputation for disinterestedness, but Necker and Washington did so because of the intrinsic satisfaction it provided.

The effects of the internal audience are, by the nature of the case, more difficult to document. One telltale sign is *the reluctance to admit a mistake*.[21] Earlier I cited La Rochefoucauld's comment on how painful it is to disapprove of what we once approved. Some scholars seem constitutionally incapable of

[21] It is of course even more difficult to admit mistakes before an external audience.

saying, "I was wrong." Another instance is summarized in Seneca's observation, "Those whom they injure, they also hate" and in a French proverb, "Who has offended can never forgive."[22] In such cases, we may observe a transmutation, a rewriting of the script, by which the victim becomes a perpetrator. A possible explanation of the sunk-cost fallacy (see Chapter 14) is that we would rather persist in a venture that has acquired negative net present value than to admit that we should not have embarked on it. Strictly speaking, of course, embarking on it may not have been an error *at the time*, but the hindsight fallacy might lead us to think that if it had negative present value today it would have had so from the outset.

Conversely, amour-propre induces a tendency to think that our choices must be good because they are *ours*, leading us to overvalue the restaurant meals we order or the friends we make. Less trivially, members or officials in an institution naturally come to think of it as an important one, and resist attempts to reduce its importance. Constituent assemblies that also exercise legislative functions may for that reason (among others) tend to give more importance in the constitution to the legislature than do assemblies that have constitution making as their only task.[23] The mixed legislative/constituent assemblies in Poland (1921) and in France (1946) created highly legislative-centric constitutions.

Other effects of amour-propre include our tendency to believe that others spend more time thinking about us than they actually do, and our tendency to believe that others will act to promote our interest rather than their own. George Eliot says about Mr. Casaubon in *Middlemarch* that he was "the centre of his own world [and] liable to think that others were providentially made for him, and especially to consider them in the light of their fitness for the author of [his magnum opus] a 'Key to all Mythologies.'" The *performance* of others may also impinge on our amour-propre, enhancing it if it is inferior to ours and threatening it if it is superior. Thus what I shall call the "first-order pain of envy" may cause us to rewrite the script in ways I discuss in the next section.

### The need to see oneself as guided by a motivation that is highly ranked in the hierarchy of motivations

People may be praised or blamed for what they do. They may also be praised and blamed for the motivations on which they act. In his essay on Coriolanus,

[22] Aristotle claimed that we love those whom we help because *we* helped them, perhaps another effect of amour-propre.

[23] Among the other reasons is that if the framers aspire to be elected to the first ordinary legislature, they might want that body to be powerful. Also, because they *know* more about the legislature than about the other branches of government, they might easily come to exaggerate its importance.

Plutarch wrote that "one great reason for the odium he incurred with the populace in the discussions about their debts was, that he trampled upon the poor, not for money's sake, but out of pride and insolence." In other words, greed may have been seen as a more acceptable motivation than pride and insolence. The distinction is reflected in the phrase, "adding insult to injury."

In any given society or subculture, there is a normative hierarchy of motivations. Some motivations are highly ranked, others are more lowly. In classical Athens, they were ranked in roughly the following order: concern for the public interest, anger, self-interest, hubris (the "infliction of dishonor for the pleasure of expressing a sense of superiority"), and envy.[24] A man might counter a charge of hubris by saying that he was drunk or angry. Also, among the Athenians, the desire to seek revenge for an insult was stronger and more highly respected than the desire to seek revenge for an injury; in modern societies, the opposite may be true. ("Sticks and stones will break my bones, but words will never harm me.") In most contemporary Western societies, anger and the desire for vengeance probably occupy a lower place in the hierarchy than the pursuit of self-interest. Tocqueville claimed that among the Americans he observed, the pursuit of private interest was ranked above public-interested motivations. This may remain true for contemporary Americans. In the United States today, therefore, motivations might perhaps be ranked as follows: self-interest, public interest, anger, and envy. In societies characterized by "amoral familism," the desire to promote the public interest might be ranked at the very bottom of the hierarchy, and the desire to take revenge for a slight, even a slight one, at the top. These impressionistic rankings may of course be questioned. What does not seem questionable, however, is the existence in all societies or groups of such hierarchies.

The normative hierarchies induce two effects, mediated by, respectively, the external and the internal audience. The first, which is easier to observe if not necessarily the most important, arises because people have an incentive to misrepresent their motivations to other people. The effect is obvious and richly documented. Here, I focus on the effect that arises when people internalize the hierarchy and *want to look good in their own eyes*. In the concocted example I presented at the beginning of the chapter, the higher status of the concern for health compared to vanity was responsible for the transmutation. I shall now offer two non-concocted examples, taken from studies of nineteenth- and twentieth-century America.

Commenting on America around 1830, but phrasing the argument in more general terms, Tocqueville wrote that:

[24] When Iago says, "If Cassio do remain, he hath a daily beauty in his life that makes me ugly," he is being unusually, perhaps implausibly frank.

In a democracy, ordinary citizens see a man step forth from their ranks and within a few years achieve wealth and power. This spectacle fills them with astonishment and envy. They try to understand how a man who only yesterday was their equal today enjoys the right to rule over them. *To attribute his rise to talent and virtue is inconvenient, because it requires them to admit that they are less virtuous and clever than he*. They therefore ascribe primary responsibility to some number of his vices, and often they have reason to do so. This results in I know not what odious mingling of the ideas of baseness and power, unworthiness and success, utility and dishonor.

The sentence I have italicized refers to the operation of amour-propre. Because the feeling of inferiority is so painful, citizens tell themselves a story to alleviate it. According to the story, the superior individual achieved his position by immoral or illegal means, and perhaps even at their expense.[25] As a by-product, the emotion of envy is replaced by indignation or anger. So far, the analysis does not refer to any normative hierarchy. I suggest, however, that an additional mechanism can be at work: because of the low status of envy in the motivational hierarchy, noted by Plutarch, Adam Smith, and no doubt many others, individuals are motivated to transmute the emotion into one that occupies a higher status. The first-order pain of envy is triggered by the thought, "I am inferior." The second-order pain is triggered by the thought, "I am envious." Thus when Morris Zapp in David Lodge's novel *Small World* receives a letter from a rival saying he has met a wonderful woman and asking him to reserve a hotel room for them, he says to himself that the "letter reads like the effusion of some infatuated teenager. Morris will not admit to himself that there may be a trace of envy in his harsh assessment. He prefers to identify his response as righteous indignation at being more or less compelled to take part in the deception" of the rival's wife.

Tocqueville's claim that the Americans he observed ranked private interest above the public interest also seems to be valid for recent times. Although American citizens spontaneously perform a number of disinterested actions, they do not always acknowledge their motivation to others or to themselves. Instead, they come up with all sorts of excuses for doing good. Thus one study found that "although people actually engage in many acts of genuine compassion, they are loathe to acknowledge that these acts may have been motivated by genuine compassion or kindness. Instead, people offer pragmatic or instrumental reasons for them, saying things such as 'It gave me something to do' or 'It got me out of the house.' Indeed, the people . . . seemed to go out of their way to stress that they were not a 'bleeding heart,' 'goody two-shoes,' or 'do-gooder.'"

The reluctance to present oneself as animated by disinterested motives might also be due, however, to the suspicion that others might disbelieve the claim. To illustrate the point, and also provide some evidence about the

[25] It has been persuasively argued that the same transmutation is at the origin of anti-Semitism.

normative hierarchy in classical Athens, let me cite a comment on the Athenian "sycophants." These were professional accusers who initiated lawsuits for private gain, either because they could hope for a share of the fine or because they hoped that even innocent plaintiffs would settle in private rather than risk litigation. As sycophants were regarded (according to Aristotle) with hatred, it was important for them to misrepresent their motivation. According to one scholar, it was more effective to disguise their interest as passion than to try to pass themselves off as motivated by impartial motives: "When a citizen appeared in court as a public accuser his first anxiety was . . . to dispel any suspicion that he was a sycophant. He could stress his public-spiritedness, but that tends to make ordinary folk even more suspicious, and usually there was a much more cogent argument to deploy: he could declare that the accused was his personal enemy and that he was using his citizen right to prosecute for revenge and not for gain."

### Self-poisoning of the mind

Many of the transmutations I have discussed make the agent *better off* in some respect. The statement by a character in Ibsen's *The Wild Duck*, "if you take the life lie from an average man, you take away his happiness as well," is not always true, but it accurately captures many cases familiar from everyday life and from world literature. Examples include getting rid of the first-order and second-order pains of envy, rewriting the script so that a failure appears to be the effect of the malign behavior of others, the retail deprecation of the wealthy and the beautiful, and the wholesale deprecation of wealth and beauty.

Yet the benign effect of illusions and delusions cannot be the whole story. Wishful thinking can make us feel good here and now, but we can fall flat on our face if we use motivated beliefs as premises for action. (Pissing in your pants to keep warm has only a momentary effect.) People who are so strongly subject to reactance that they *never* accept doing what others suggest to them may miss out on many opportunities. As noted earlier in the chapter, the search for novelty may cause people to "improve themselves to ruin." In Chapter 4, I observed that cognitive dissonance reduction can induce preference changes that make the agent worse off, as judged by the pre-reduction preferences.

I shall now present two episodes from Proust. In both, we observe the transmutation of "I cannot have it" into "I do not want it," but with opposite effects on the happiness of the agents. The first occurs when the Narrator observes the behavior of two bourgeois wives toward an old, rich, and titled lady who stays at the same hotel:

> Whenever the wives of the notary and the magistrate saw her in the dining-room at meal-times they put up their glasses and gave her an insolent scrutiny, as minute and distrustful as if she had been some dish with a pretentious name but a suspicious

appearance which, after the negative result of a systematic study, must be sent away with a lofty wave of the hand and a grimace of disgust. No doubt by this behavior they meant only to show that, if there were things in the world which they themselves lacked – in this instance, certain prerogatives which the old lady enjoyed, and the privilege of her acquaintance – it was not because they could not, but because they did not want to acquire them. But they had succeeded in convincing themselves that this really was what they felt; and it is the suppression of all desire for, of all curiosity as to forms of life which are unfamiliar, of all hope of pleasing new people (for which, in the women, had been substituted a feigned contempt, an artificial cheerfulness) that had the awkward result of obliging them to label their discontent satisfaction, and lie everlastingly to themselves, two conditions for their being unhappy . . . The atmosphere of the microcosm in which the old lady isolated herself was not poisoned with virulent bitterness, as was that of the group in which the wives of the notary and magistrate sat chattering with impotent rage.

The second episode involves the absurdly self-contented father of the Narrator's friend Bloch.

[He] lived in the world of approximations, where people salute the empty air and arrive at wrong judgments. Inexactitude, incompetence do not modify their assurance; quite the contrary. It is the propitious miracle of amour-propre that, since few of us are in a position to enjoy the society of distinguished people, or to form intellectual friendships, those to whom they are denied still believe themselves to be the best endowed of men, because the optics of our social tiers make every grade of society seem the best to him who occupies it, and beholds as less favored than himself, less fortunate and therefore to be pitied, the greater men whom he names and calumniates without knowing, judges and despises without understanding them. Even in cases where the multiplication of his modest personal advantages by his amour-propre would not suffice to assure a man the dose of happiness, superior to that accorded to others, which is essential to him, envy is always there to make up the balance. It is true that if envy finds expression in scornful phrases, we must translate "I have no wish to know him" by "I have no means of knowing him." That is the intellectual sense. But the emotional sense is indeed, 'I have no wish to know him'. The speaker knows that it is not true, but he does not, all the same, say it simply to deceive; he says it because it is what he feels, and that is sufficient to bridge the gulf between them, that is to say to make him happy.

Although the reference to envy is hard to understand,[26] the overall idea is clear: the upgrading of one's own small advantages may, if necessary, be supplemented by the downgrading of the greater advantages of others, to produce happiness.

The bourgeois wives are subject to the *ressentiment* that Max Scheler, inspired by Nietzsche, called a "self-poisoning of the mind." Scheler claimed,

[26] Proust's reference to the downgrading tendency of envy is distinctly idiosyncratic. Envy presupposes the recognition of the value of the envied object, not the denial of its value. The action tendency of envy is to destroy what you cannot get, not to denigrate it. Other passages show that Proust was perfectly aware of this standard understanding.

though, that the reaction is "chiefly confined to those [who] fruitlessly resent the sting of authority." The passage from Proust suggests that the idea has broader application.

### Bibliographical note

The present chapter draws heavily on Chapter V of my *Alchemies of the Mind* (Cambridge University Press, 1999) and, especially, on Chapter 3 of *L'irrationalité* (Paris: Seuil, 2010). The article by Otto Neurath, "The lost wanderers of Descartes and the auxiliary motive," is reproduced as Chapter 1 of his *Philosophical Papers 1913–1946* (Dordrecht: Reidel, 1983). The claim that an agent may reduce his dissonance by first increasing it is in L. Festinger, *A Theory of Cognitive Dissonance* (Stanford University Press, 1957), p. 129. I discuss Proust's examples of transmutation at greater length in Chapter 15 of *L'irrationalité* and in "Self-poisoning of the mind," *Philosophical Transactions of the Royal Society B* 365 (1010), 221–26. The observation on Stendhal's unhappy love affair is in M. Crouzet, *Stendhal* (Paris: Flammarion, 1990), p. 292. Studies of conspiracy theories in France include G. Lefebvre, *La grande peur de 1789* (Paris: Armand Colin, 1988), S. Kaplan, "The famine plot persuasion in eighteenth-century France," *Transactions of the American Philosophical Society* 72 (1982), 1–79, and F. Ploux, *De bouche à oreille: naissance et propagation des rumeurs dans la France du XIXe siècle* (Paris: Aubier, 2003). My comments on agency bias in the explanation of cancer are based on G. Johnson, *The Cancer Chronicles* (New York: Knopf, 2013) and on D. Holzman, "President's cancer panel stirs up environmental health community," *Journal of the National Cancer Institute* 102 (2010), 1106–13. The quotation from William Buckley is in R. Perlstein, *Before the Storm* (New York: Nation Books, 2001), p. 155. The passage on the human mind's propensity to teleological reasoning is taken from the editorial introduction to A. Vayda and B. Walters (eds.), *Causal Explanation for Social Scientists* (Lanham, MD: Altamira, 2011). The comment on the pleasure and comfort provided by analogies are from J. H. M. Salmon, *Society in Crisis* (London: Routledge, 1979), p. 14 and from a review by Bernard Williams of a book by Lucien Goldmann, reprinted in his *Essays and Reviews 1959–2002* (Princeton University Press, 2014), p. 80. The discussion of analogies in debates over the Vietnam War owes much to Y. Khong, *Analogies at War: Korea, Munich, Dien Bien Phu, and the Vietnam Decisions of 1965* (Princeton University Press, 1992). Rostow's analogy between the Vietnam War and World War II is in D. Milne, *America's Rasputin: Walt Rostow and the Vietnam War* (New York: Hill and Wang, 2008), pp. 170–1. The comment on McNamara's limited experience is in H. McMaster, *Dereliction of Duty* (New York: Harper, 1997), p. 328. The reference to the reactance of doctors is in R. Dijkstra *et al*.,

"Perceived barriers to implementation of diabetes guidelines in the Netherlands," *Netherlands Journal of Medicine* 56 (2000), 80–5. The findings about reactance induced by suggested taxi tips are in K. Haggag and G. Paci, "Default tips," working paper (Department of Economics, Columbia University, 2013). The biography of Mitterrand is by Philip Short, *Mitterrand: A Study in Ambiguity* (London: The Bodley Head, 2013). An envy-based explanation of anti-Semitism is in B. Netanyahu, *The Origins of the Inquisition* (New York: Random House, 1995), p. 989. The passage on Athenian sycophants is from M. H. Hansen, *The Athenian Democracy in the Age of Demosthenes* (Oxford: Blackwell, 1991), p. 195. The definition of hubris is from N. R. E. Fisher, *Hybris* (Warminster: Aris and Phillips, 1992). The passage asserting that people may be loathe to acknowledge that their acts are motivated by kindness is in R. Wuthnow, *Acts of Compassion* (Princeton University Press, 1991), cited from D. Miller, "The norm of self-interest," *American Psychologist* 54 (1999), 1053–60.

*Part III*

# Action

Although I shall largely use "action," "behavior," "decision," and "choice" as synonymous terms, it can be useful to distinguish among them. The broadest category is *behavior*, understood as any bodily movement whose origin is internal to the agent, not external (as when he is carried away by a landslide). *Action* is intentional behavior, caused by the desires and beliefs of the agent. Thus reflex behaviors are not actions; having an erection is not an action (but it may be induced by one, such as taking Viagra); falling asleep is not an action (but may be induced by taking a sleeping pill).

An action may or may not be preceded by a conscious *decision*. When Pascal said that we are "automata as well as minds" and Leibniz that "we are empirical in three-quarters of our actions," they referred to the habitual and unthinking character of much everyday behavior. When I drive to work along my usual route I do not consciously decide to turn right here and left there. The very first time or times I drove to work, however, the actions were preceded by explicit decisions. In fact, they were preceded by an explicit *choice* among alternative paths. Although all choices are decisions, the converse is not true. When I decide to pick up the book I have been reading, I need not have any explicit alternative in mind. I see the book on the table; the sight reminds me that I have enjoyed reading it; and I decide to pick it up. No choice is involved.

The focus in Part III is on choice. I believe that the concept of choice is the most fundamental idea in the social sciences. In Part II, I considered the subjective precursors of choice: preferences (desires, motivations), beliefs, emotions, and prejudices, as well as some precursors of these precursors. In the following chapters, I consider the mechanisms by which these precursors generate choices and, usually, actions. I say "usually," because *not all choices lead to actions*. One may choose not to do something, for example, not to save a drowning person if the intervention would be at some risk to oneself. If the person drowns and no third parties are involved, I have no causal responsibility for the outcome. I may have a moral and, in some countries, a legal responsibility, but that is another matter.[1]

[1] In the United States, there is no duty to be a Good Samaritan, except under narrowly circumscribed conditions. In continental Europe, "non-assistance to a person in danger" can be severely

But suppose there is a third party present or, as in the "Kitty Genovese" case, many parties. If the third party observes that I am in a position to help the drowning person and that I do not, he or she might reasonably draw the inference that the situation is less serious than it would otherwise have seemed and, as a result, also abstain from helping. In that case, my choice to do nothing would have caused another person to choose to do nothing. Thus choices can have causal efficacy even when they do not generate an action.

*Constraints* and *selection* offer two alternative paths to the explanation of behavior. In Chapters 10 and 11 I consider the explanatory power of these two objective factors. When subjective factors are constant over time or invariant across agents, changes or differences in the constraints they face may explain changes or differences in behavior. More generally, subjective factors and constraints interact, in complex ways, to produce actions. Selection mechanisms rely on (mostly) random variation and (mostly) deterministic selection among the randomly generated variants, without the intervention of choice. Yet in some cases the sources of variation and the mechanisms of selection also include choice. Whereas selection is hugely important in biology, I believe it is marginal in the social sciences. Because humans are animals, some basic action tendencies are shaped by natural selection, but with few exceptions they can be modulated or suppressed by individual or collective choice.

Chapters 13 through 15 are organized around rational-choice theory and its alternatives. In these chapters, I deal with individual actions only. In Chapters 18 and 19, I consider rational-choice explanations in interactions among several agents. As I discuss in Chapter 14 and in the Conclusion, I have come to be more skeptical of rational-choice explanations of action than I used to be. Yet although much behavior is irrational in one way or another, there is a sense in which rationality remains primary. Human beings *want* to be rational. We do not take pride in our lapses from rationality. Rather, we try to avoid them or correct them, unless our amour-propre prevents us from recognizing them (Chapter 9).

I conclude Part III by considering the production of works of art, mainly novels and plays, as *actions* that can be explained along the same lines as other actions. (Some observations on this topic are also offered in Chapter 10.) The advantage of this approach is that there is a *fact of the matter* by virtue of which a given explanation, if right, is right and, if wrong, is wrong.

punished if the risk to the Samaritan of helping is small compared to the danger of the person in need of help. Some American legal scholars argue that the American law is more efficient, since a general duty to assist would create an incentive for potential rescuers to avoid locations where rescues are likely to be needed, because of the threat of liability. This may or (more likely) may not be so; what does seem certain is that the American system did not come into being for efficiency reasons, nor did it come into being for other reasons and then continue in effect because of its efficiency properties.

# 10 Constraints: opportunities and abilities

To characterize and explain behavior, we sometimes say, "He did the best he could." Here, "the best" is defined by the agent's desires or preferences. What the agent "could do" is defined by his *opportunities* and his *abilities*. An opportunity, according to the Oxford English Dictionary, is "a time, condition, or set of circumstances permitting or favorable to a particular action or purpose." An ability is the "quality in a person or thing which makes an action possible; suitable or sufficient power or proficiency." While not unambiguous, the definitions suggest that an opportunity is an enabling factor external to the agent, whereas an ability is an enabling factor internal to him or her.

Typically, an agent needs both opportunities and abilities to realize her desires. If I have a horse and need to get somewhere in hurry, but do not know how to ride, I am likely be thrown off. If I have the skill and the desire, but lack a horse, I will not be going anywhere. ("A horse, a horse, my kingdom for a horse!") A basketball player will not score if others do not provide him with an opportunity, by passing the ball to him, or if he is too unskilled to profit from his opportunities.

In military parlance, the idea of *capabilities* usually includes both opportunities and abilities, typically related to action in a multiplicative rather than additive fashion. A commander has to estimate both the weaponry available to the enemy and the skill to use it. Advanced materiel in the hands of soldiers not trained to use it may be worthless. Some philosophers argue that capabilities – defined as "a person's opportunity and ability to generate valuable outcomes" (Wikipedia) – should be the central concept in a theory of distributive justice. For some purposes, capabilities may indeed be the most relevant factor. For my purposes, I want to consider the two components of capabilities separately.

In most of the cases I shall discuss, desires, opportunities, and abilities are jointly sufficient and individually necessary to produce an action. If we observe that an agent does *not* perform some action and we want to know why (see Chapter 1 on "why-explanations" of non-events), the absence of any one of the three factors might provide the answer. Citing the absence of more than one factor would provide an overdetermined explanation. In *Democracy*

*in America,* Tocqueville provides a striking example when he argues that all *three* factors were absent. As other observers have been, he was puzzled by what he perceived as the low intellectual and moral quality of elected politicians in the United States. He argued that American citizens had neither the *opportunity* to select good leaders (because qualified persons do not to stand for office), nor a *desire* to do so (because of envy of superior individuals), nor the *ability* (because of lack of cognitive discernment).[1] In other cases, *two* factors may be absent. As I discuss later, both Tocqueville and James Madison cite the absence of two factors – desires and opportunities – to account for certain non-events (see Chapter 1).

In standard cases, desires, opportunities, and abilities are given independently of each other. Sometimes, however, the factors may be related through a common cause, or one of them may directly influence another. I shall give examples of both cases.

### Desires and opportunities

Opportunities are the options or means available to the agent. If we ask, "Why did he do it?," the answer "Because it was his best option" provides a rudimentary rational-choice account of the behavior. In many cases, more is needed to provide a satisfactory rational-choice explanation. These complications will concern us in Chapter 13. Here, I discuss how far the simple desire–opportunity framework can get us, and some limitations of that framework. In particular, the usually tacit assumption that agents are *aware* of the opportunities open to them is not always justified.

There is another, equivalent way of looking at the matter. In understanding behavior, we may begin with all the abstractly possible actions the individual might undertake. The action that we actually observe can be seen as the result of two successive filtering operations. The first filter is made up of all the *opportunity constraints* – physical, economic, legal, and others – that the agent faces. The actions consistent with all these constraints constitute the opportunity set.[2] The second filter is a mechanism that determines which action within the opportunity set will actually be carried out. Here, I am assuming that the agent chooses the action that will have the best

[1] The factors may not be independent of each other. If the voters wanted good leaders and had the competence to identify them, more might have been forthcoming. Later I cite several interaction effects of this kind. Note also that the second and third explanation are in some tension: to reject outstanding individuals out of envy, citizens must be able to identify them.

[2] When I refer to legal constraints, I do not mean the effect of laws in making certain actions more costly than others (that belongs to the second filter), but to their effect in making them possible or impossible. I cannot vote at times other than election days or get married outside the legally determined venues.

consequences, as assessed by his desires (or preferences). In later chapters, we shall consider other second-filter mechanisms.

The filter approach suggests the following question: what if the constraints are so strong that there is nothing for the second filter to work on? Can it happen that the constraints uniquely determine one and only one action that is consistent with all of them? The rich and the poor alike have the opportunity to sleep under the bridges in Paris, but the poor may have no other opportunity.[3] For a poor consumer, economic and calorific constraints might jointly determine a unique bundle of goods.[4] Those who defend the idea of *structuralism* in the social sciences may be understood as saying that constraints typically are so strong as to leave very little or no scope for choice.[5] *Why* this should be so, however, remains mysterious. One cannot argue, for example, that the rich and powerful make sure that the poor and oppressed have no other option than to work for them, since this statement presupposes that the rich and powerful, at least, do have a choice (see next paragraph).

The constraints may also be so strong that *no* action satisfies them all. In that case, the constraints may provide a why-explanation of a non-event. Time constraints and space constraints, for instance, may jointly preclude action. In mid-nineteenth-century Massachusetts, "polls were closed at sunset, to disfranchise, according to Democrats, Cambridge laborers who worked in Boston but had to vote in their place of residence. When they returned home the polls were closed." Note that the explanation also cites the *choice* of the Republicans to prevent Democrats from voting.

Even when behavior is the joint result of desires and opportunities, *variation* in behavior over time may be largely explained by opportunities. Alcohol consumption is, in general, determined both by the strength of people's desire to drink, compared to their other desires, and by what they can afford. When alcohol prices rise steeply, for instance, in wartime, consumption falls sharply.

One explanation could be in terms of indifference curves (Figure 10.1). Suppose that the consumer has to allocate her income between alcohol and some bundle of ordinary consumption goods. The relative prices and her income are initially such that she faces the opportunity set inside the triangle OAA′. Assuming she spends all her income, we can limit ourselves to the

[3] If the opportunity set is described in finer grain, the poor may have the choice as to which bridge to sleep under. This is a general point: for any description of the options one may be able to specify a situation in which the agent has only one feasible option, but for any situation one may find a description under which there is more than one.

[4] Typically, however, there are several strategies for surviving at subsistence level. Unlike the choice of which bridge to sleep under, these often differ in non-trivial ways. One of the flaws in Marx's labor theory of value stems from his failure to understand this fact.

[5] Another, unrelated idea of structuralism was considered in Chapter 1.

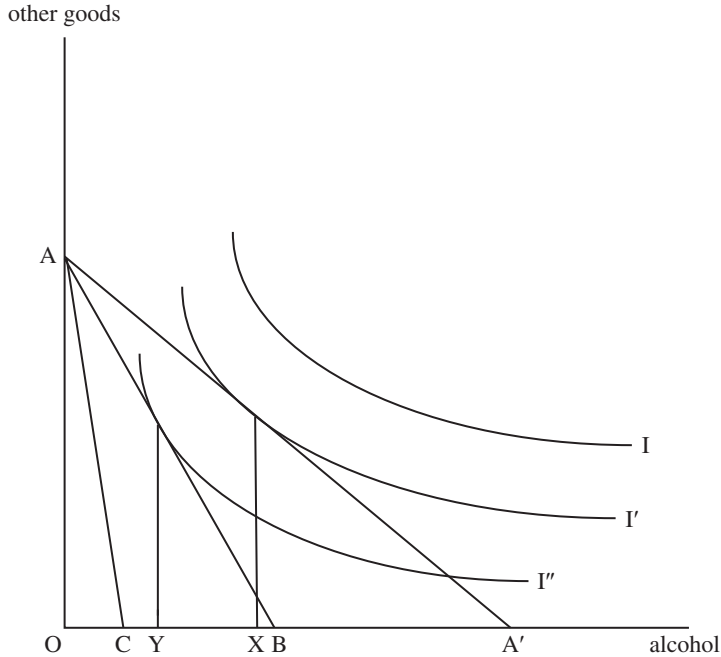



Figure 10.1

budget line AA′.[6] The strength of her desires for alcohol versus the consumption bundle is shown in the shape of the *indifference curves* I, I′, and I″. The term reflects the idea that the consumer is indifferent among all the combinations of alcohol and other goods that lie on any given curve, while preferring any combination on a higher curve to any on a lower curve.[7] To choose the best among the options available to her, the consumer must pick the point on the budget line that is tangent to an indifference curve, since this will be the highest among the curves that include a combination she can afford. In Figure 10.1, this yields alcohol consumption OX.

Suppose now that the price of alcohol goes up so that the consumer faces the budget line AB. As the point of tangency has moved to the left, the consumer will now consume OY. We could carry through the same reasoning if a further price increase shifts the budget line to AC. Yet even if we know nothing about

[6] I ignore, in other words, that the person might work overtime, make her own liquor, or buy smuggled goods. In policy applications, these issues are important.

[7] The shape of the curves corresponds to the fact that the more alcohol the agent is currently consuming, the more alcohol she needs to compensate (to remain at the same level of welfare) for a given cut in the consumption bundle.

the shape of the indifference curves, we can predict that in this situation the consumer will not consume more than OC, which would be the case if she spent all her income on alcohol. The opportunity set by itself can explain a great deal of the variation over time. The second filter, in fact, could be anything – optimizing behavior, an irresistible craving for alcohol, custom, or whatever – the consumer would still be severely restricted by the first filter.

I chose this particular example to discuss the question of allegedly "irresistible" desires, such as the desire of drug addicts, heavy smokers, or alcoholics for the substance to which they are addicted. Are drugs more like insulin, which a diabetic will buy at any price, or more like sugar, of which consumers routinely purchase less when the price goes up? As proof that they are more like sugar, one often cites the fact that drug consumption goes down when prices go up. Yet as we have seen, that might simply be due to the inability of the addict to consume beyond his budget. (The diabetic, too, might be unable to purchase the insulin he needs if the price goes up.) Thus it seems that the fall of alcohol consumption during wartime is often due to its unavailability, leaving the question of irresistible versus resistible desires unresolved. We do know on other grounds, however, that alcohol consumption *is* price sensitive. Even when consumers can afford to maintain their previous level of consumption at higher prices, they do not.

One can also argue that opportunities are more basic in a further respect: they are easier to observe, not only by social scientists, but also by other individuals in society. In military strategy, a basic dictum is that one should plan on the basis of the opponent's (verifiable) opportunities, not on his (unverifiable) intentions.[8] If we have reason to believe that the opponent *might* have hostile intentions, the dictum can lead to worst-case assumptions: the opponent will hurt us if he can. The situation is complicated by the fact that our belief in the hostile intentions of the opponent may be grounded in a perception that *he* believes we have the means and perhaps the intention of hurting him. In this morass of subjectivity, objective opportunities may seem to provide the only firm basis for planning. In strategic debates over the Vietnam War, Lieutenant General Goodpaster protested to Secretary McNamara in the fall of 1964, "Sir, you are trying to program the enemy and that is one thing we must never try to do. We can't do his thinking for him."

[8] Overestimations of Soviet and Iraqi military might remind us, however, that even opportunities can be hard to verify. Just as agents may have an incentive to misrepresent their intentions and preferences, they may have an incentive to misrepresent their opportunities, by making an opponent believe that they have more or fewer means at their disposal than they really have. In Roman armies, "a common military stratagem was to enlarge or reduce the size of the camp in order to intimidate the enemy with the prospect of a larger force or to encourage the enemy to underestimate the force."

Still another reason why opportunities may appear more fundamental than desires has to do with the possibility of influencing behavior. It is usually easier to change people's circumstances and opportunities than to change their minds.[9] This is a cost–benefit argument about the dollar effectiveness of alternative policies – not an argument about relative explanatory power. Even if the government has a good theory, allowing for explanation as well as prediction, it may not allow for much control, since the elements on which it can act may not be the causally important ones. Suppose that the government is persuaded that weak economic performance can be traced back to risk-averse businesspeople and to strong unions. It may be fully convinced that the mental attitude of managers is the more important cause yet be unable to do anything about it. By contrast, as the Reagan and Thatcher years showed, unions can be broken by government action.

For an important example, consider suicidal behavior. To commit suicide, the desire to kill oneself is not enough: one must also find the means to do it. The high suicide rate among doctors, for instance, may be due in part to their easy access to lethal drugs, which are the favored means of suicide in this group.[10] Although the government may try to limit suicidal intentions, by providing help lines or persuading the media to play down the reporting of suicides, which can trigger suicide by contagion, the most effective results are obtained by making access to the means of suicide more difficult.[11] Policies include barriers that make it more difficult to jump from bridges or tall buildings, more rigorous control of certain prescription drugs, restrictions on the sale of handguns, the replacement of lethal carbon monoxide by natural gas in kitchen ovens, and the installation of catalytic converters that reduce the carbon monoxide emissions in motor vehicle exhausts. In the future, we may see the banning of "suicide help" internet sites. (Strictly speaking, this measure would not eliminate any opportunities, only knowledge about them.) The simple switch from bottles to blister packs has contributed to the reduction of the number of suicides and severe liver damage from paracetamol poisoning. Although strong, the urge to kill oneself is so short-lived that by the time one has managed to open all the blisters it may have subsided. Reducing the maximum number of tablets that can be available in individual preparations has also reduced the likelihood of severe poisonings. By the time one has done the round of pharmacies to buy enough bottles, the urge may

[9] In addition, as argued later, the best way to change their minds may be to change their circumstances.

[10] Surprisingly, police officers' easy access to guns does not make them more suicide prone than others.

[11] When suicide rates fell radically in Britain in the 1970s, the change was initially attributed to the helplines established by the Samaritan Centres but later explained by the shift from lethal coal gas to the less lethal natural gas in domestic ovens.

have subsided.[12] In China, interrogations of officials suspected of corruption are held in rooms at the ground level, to prevent them from killing themselves by jumping out of the window (*Financial Times*, October 17, 2014).

To be sure, a determined individual will usually find a way. When one common means of taking one's life is removed, the ensuing drop in the suicide rate may to some extent be a temporary one. Yet in some cases at least, the effect seems to have been lasting, as one would expect. If the urge to kill oneself is fleeting rather than firmly anchored, it might be gone by the time one manages to get hold of a suitable means.[13] Hence merely *delaying* (rather than blocking) access to means could be effective in preventing impulsive suicides. The requirement of a waiting time before the purchase of a handgun could reduce suicide as well as homicide rates.[14] I pursue this issue in Chapter 15.

A more complex example of desire–opportunity interactions may be taken from Madison's analysis of factions in *The Federalist* # 10. He argues that to prevent a factious majority from oppressing the minority, "Either the existence of the same passion or interest in a majority at the same time must be prevented, or the majority, having such coexistent passion or interest, must be rendered, by their number and local situation, unable to concert and carry into effect schemes of oppression. If the impulse and the opportunity be suffered to coincide, we well know that neither moral nor religious motives can be relied on as an adequate control." Objectively, the members of a factious majority have the opportunity to oppress the minority. Yet, as Madison goes on to argue, concerted action may be difficult if they do not *know* they have that opportunity. If you extend the size of the republic, "you make it less probable that a majority of the whole will have a common motive to invade the rights of other citizens; or if such a common motive exists, it will be *more difficult for all who feel it to discover their own strength*, and to act in unison with each other."[15] Madison's argument is, as it were, double-barreled: not only will a large republic prevent factious majorities from arising, but it will also prevent concerted action were one such majority to arise. Conversely, in

[12] "In one study of people who survived a suicide attempt, almost half reported that the whole process, from the first suicidal thought to the final act, took 10 minutes or less" (*New York Times*, March 10, 2015).

[13] I do not think the increased *cost* of locating a means could deter suicide. There may be cost–benefit considerations involved in a decision to kill oneself, such as weighing the pain the agent will inflict on others against his or her own relief from pain, but the cost of finding an appropriate means will not, for a determined individual, make a difference.

[14] As a matter of fact, most of the American states that impose a cooling-off period before the purchase of a handgun do so to give the authorities time to see whether the prospective buyer has a criminal record or a history of mental illness, not to create a cooling-down period for the buyer.

[15] The argument I have italicized refers to what later came to be called pluralistic ignorance (Chapter 22).

small republics such majorities are not only more likely to arise, but also more likely to organize themselves.

In Tocqueville's *Democracy in America* we find a large number of arguments with this "not only" structure. As an example, consider his discussion of the impact of slavery on slaveowners. In the first place, slavery is unprofitable, compared to free labor. "The free worker receives wages, the slave receives an upbringing, food, medicine, and clothes; the master spends his money little by little in small sums to support the slave; he scarcely notices. The workman's wages are paid all at once and seem only to enrich the man who receives them; but in fact the slave has cost more than the free man, and his labor is less productive."[16] But "the influence of slavery extends even further, penetrating the master's soul and giving a particular turn to his ideas and tastes." Because work is associated with slavery, the southern whites scorn "not only work itself but also enterprises in which work is necessary to success." They lack both the opportunities and the desire to get rich: "Slavery ... not only prevents the white men from making their fortunes but even diverts them from wishing to do so." If Tocqueville is right, a classic debate over the economic stagnation of slave societies is spurious. There is no need to ask whether lack of investment desires or lack of investment opportunities provides the correct explanation: both sides could be right.

In other contexts, Tocqueville argues that in some societies the lack of desire may provide why-explanations of non-events that, in others, are explained by lack of opportunities. Comparing ancient and modern (American) slaves, he writes that in "Antiquity, people sought to prevent the slave from breaking his chains; nowadays they seek to sap his desire to do so ... It was noticed long ago, moreover, that the presence of the freed Negro was vaguely disquieting to the souls of the unfree, in whom it aroused the first glimmering of an idea of their rights. In most cases, therefore, Americans in the South stripped masters of their prerogative to free their slaves." In a striking anticipation of Tocqueville's argument, a Virginia politician said that "If the blacks see all of their color slaves, it will seem to them a disposition of Providence and they will be content." For the same reason, many whites wanted free blacks to be deported to an African colony, where they would be out of sight and out of the mind of the slaves.

The argument has a wide application. Writing in 1800, Governor Monroe of Virginia was puzzled by the slave insurrection that took place that year. "It seemed strange that the slaves should embark in this novel and unexampled enterprise of their own accord, [as] their treatment has been more favorable since the revolution." He concluded that external agitators must have been at

[16] This argument is uncharacteristically opaque. A simpler argument is that except for certain branches of agriculture slavery is unprofitable because it creates no incentive for slaves to apply themselves to their work.

work. In his book on the *ancien régime*, Tocqueville solved the puzzle: "The evil one endures patiently because it seems inevitable becomes unbearable the moment its elimination becomes conceivable. Then, every abuse that is eliminated seems only to reveal the others that remain, and makes their sting that much more painful. The ill has diminished, to be sure, but sensitivity to it has increased." Roughly a hundred years later, The *Wall Street Journal* was still puzzled, when it wrote (July 18, 1966) that it "is surprising, although to an extent understandable, that the more civil rights is piled onto the statute books, the more Federal money poured into attempts at Negro betterment – the more the anger rises."[17] It is possible, though, that the 1800 and 1966 examples reflect the tendency for desires to rise faster than the means of satisfying them. Although this idea is often attributed to Tocqueville, it is not what he argued.

### Relations between desires and opportunities

Some of the arguments made by Madison and Tocqueville have a common structure: one and the same third variable shapes both desires and opportunities, which jointly shape action (or prevent it, as the case may be). In the abstract, there are four possibilities (plus and minus signs indicate positive and negative causal effects) (see Figure 10.2).

Case (A) is illustrated by Madison's analysis of direct democracies or small republics.

Case (B) is exemplified by his argument in favor of large republics and by Tocqueville's analysis of the effects of slavery on the slaveowner.

Case (C) is observed in the many cases in which lack of resources has the dual effect of increasing the incentive to improve one's situation and of reducing one's opportunity to do so. Although it is said that "necessity is the mother of invention," that is true only to the extent that hardship increases the motivation to innovate. But since innovation often requires resources (which might therefore be called "the father of invention"), the motivation by itself may not lead anywhere. Innovation often requires costly investments with an uncertain and delayed pay-off – but this is exactly what firms on the brink of bankruptcy cannot afford. Prosperous firms can afford to innovate but may not bother to do so. As the economist John Hicks said, "The greatest of all monopoly profits is a quiet life."

Similarly, while the desire to emigrate is enhanced by poverty in one's home country, the same poverty may block access to the means of emigration because of the costs of travel. Until the early nineteenth century, emigrants

[17] In some cases, reforms create demands for more reforms because they are seen as a sign of weakness on the part of the government. This mechanism differs from, but may coexist with, the tendency for reforms to make the unreformed parts of the system more intolerable.

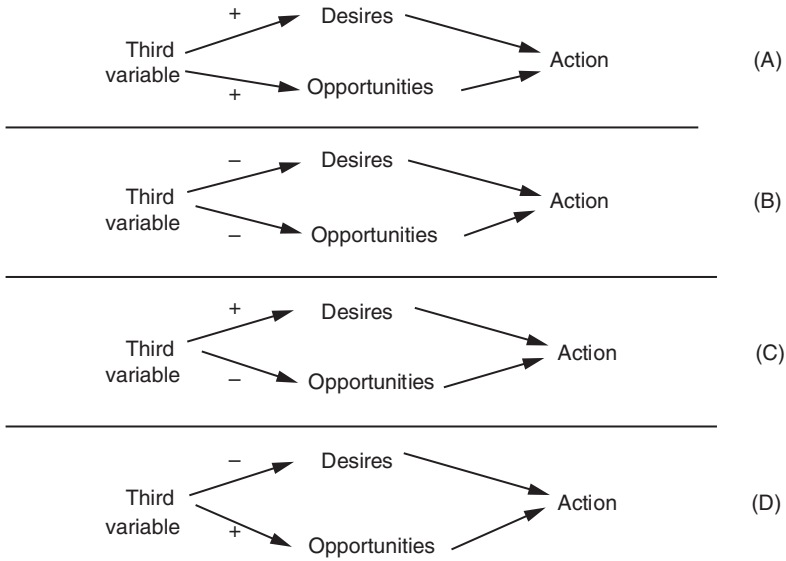



Figure 10.2

to the United States could use their bodies as collateral. Their future employers would pay for the trip in exchange for a period of indentured servitude. Today, smugglers of humans can rely on fear of the Immigration and Naturalization Service to prevent illegal immigrants from reneging on their promise to repay travel costs out of their income from labor in the receiving country. But when the Irish fled their famine in the 1840s, the poorest stayed home to die.

A further instance of case (C) is found in the study of peasant rebellions: although the poor peasants have the greatest incentive to rebel, they may not have the resources to do so. Participation in collective action requires the ability to take time off from productive activities, which is precisely what the impoverished peasant cannot afford. The middle peasants who have managed to save a bit can afford to join a rebellion, but their motivation is less acute. Marx argued that civilization arose in the temperate zones because only there did the desire for improvement meet with opportunities for improvement. Where nature is too lavish there is no desire, and where it is too scanty there are no opportunities. There may be a range of resources within which both desires and opportunities are sufficiently developed to generate action, but a priori nothing can be said about how wide or narrow it is, or whether it even exists.

We have seen an instance of case (D) in Chapter 2. The upper part of Figure 2.1 shows how Tocqueville argued that democracy (by the intermediary

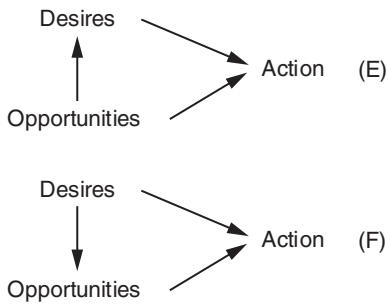



Figure 10.3

of religion) inhibited the desire to engage in the disorderly behavior for which democratic institutions such as freedom of the press and freedom of association provided an opportunity. A more commonplace observation by Tocqueville relies on the conjunction of (C) and (D), with young or old age as the third variable: "In America most rich men began by being poor; almost all men of leisure were busy in their youth; as a result, at the age when one might have a taste for study, one has not the time; and when time is available, the taste has gone."

Desires and opportunities may also affect each other directly: consider first case (E) in Figure 10.3. In Chapter 2 I touched on some of the ways in which opportunities can affect desires: people may end up desiring most what they can get or prefer what they have to what they do not. Again we may quote Tocqueville on slavery: "Is it a blessing of God, or a last malediction, this disposition of the soul that gives men a sort of depraved taste for the cause of their afflictions?" This mechanism suggests a further reason for thinking opportunities more basic than preferences. Opportunities and desires jointly are the proximate causes of action, but at the same time desires are partly shaped by opportunities through the mechanism of adaptive preference formation.

One might ask, though, whether this mechanism *matters* for behavior, since, by definition, options that are not in the opportunity set will not be chosen. Suppose the agent initially ranks options in the order A, B, C, D and then learns that A is unavailable. By adaptive preference formation, she now ranks them in the order B, A, C, D. She will choose B, as she would have had her preferences remained the same. Suppose, however, that the new ranking is C, B, A, D, inducing the choice of C. This might occur through a process of "overadaptation" to the limited opportunities. Tocqueville claimed this was a peculiar characteristic of the Frenchman: "He goes beyond the spirit of servitude as soon as he has entered it." More likely, we are dealing with a general tendency observed in many status societies. Also, the new preference ranking

might be B, C, D, A. If beautiful women reject my advances, I may console myself by the thought that by virtue of their narcissism they are actually the least desirable partners (see the previous chapter). Although my choice of partner may be unaffected, my behavior toward beautiful women in general may change (see the comments on self-poisoning of the mind in Chapter 9).

Consider finally case (F), in which the opportunity set is shaped (specifically: limited) by the agent's desire. I discuss many examples in Chapter 15, where I consider how agents can commit themselves ahead of time by deliberately limiting their future opportunities for choice. In those cases, they do so to preempt or prevent irrational or emotional decisions. *Artists* may limit their choice set for other reasons. Montaigne wrote that "just as the voice of the trumpet rings out clearer and stronger for being forced through a narrow tube so too a saying leaps forth much more vigorously when compressed into the rhythms of poetry." Proust, commenting on how his mother was forced to interrupt a conversation, writes that "she managed to extract from this constraint itself a further refinement of thought, as great poets do when the tyranny of rhyme forces them into the discovery of their finest lines."

Kant formulated the idea quite generally:

> It is advisable ... to remind ourselves that in all the free arts there is yet a need for something in the order of a constraint (*etwas zwangsmässiges*), or, as it is called, a mechanism. Without this the spirit which in art must be free and which alone animates the work, would have no body at all and would evaporate completely. This reminder is needed (in poetry, for example, correctness and richness of language, as well as prosody and meter) because some of the more recent educators believe that they promote a free art best if they remove all constraint (*Zwang*) from it and convert it from labor into mere play.

Perhaps Marx had the last sentence in mind when he made the following polemical remark:

> In the sweat of thy brow shalt thou labor! was Jehovah's curse on Adam. And this is labor for Smith, a curse. "Tranquility" appears as the adequate state, as identical with "freedom" and "happiness." It seems quite far from [Adam Smith's] mind that the individual, "in his normal state of health, strength, activity, skill, facility," also needs a normal portion of work and the suspension of tranquility. Certainly, labor obtains its measure from the outside, through the aim to be attained and the obstacles to be overcome in attaining it. But Smith has no inkling that this overcoming of obstacles is in itself a liberating activity. [Labor] becomes attractive work, the individual's self-realization [*Selbstverwirklichung*], which in no way means that it becomes mere fun, mere amusement, as Fourier with grisette-like naiveté, conceives it. Really free working e.g. composing, is at the same time precisely the most damned seriousness, the most intense exertion.

I assume that the constraints or obstacles are *chosen* by the artist, not imposed by the medium, as in silent movies before 1927, or by a principal, as when

Stravinsky, asked to write music for a ballet, replied: "I accept: how many minutes?" Imposed constraints may also liberate creativity, but do not illustrate the desire–opportunity interaction that is my topic here.

These observations provide a natural transition from opportunities to abilities. In the arts, ability cannot thrive without opportunity constraints; conversely, overcoming these constraints requires ability.

### Abilities

Ability and desire may be sufficient to generate action, if no external enabling factors are needed. A singer who desires to hit the high C needs only the ability to do so. I shall disregard such cases, and limit myself to cases where ability and desire interact with opportunities.

Earlier, I noted that ability can act as a multiplier on given opportunities. On ten opportunities to score, a good basketball player may succeed on six and an average player only on four. In addition, ability may act as a multiplier on the number of opportunities. Compared to a baseline of average ability, the multiplier may be greater than 1 if teammates pass the ball more frequently to a good player, providing him with more opportunities to score. It may also be less than 1, if the opposing team follows the good player so closely that he receives fewer passes than average players do. If both effects occur, the net effect on the opportunities can be indeterminate. The possibilities are shown in Figure 10.4.

Even if the net effect on *opportunities* is negative, however, the net effect on *scores* may be positive, if the greater ability to score more than compensates for the smaller number of opportunities to score. Moreover, even if the net effect on *his* scores is negative, his *team* may benefit if the players assigned to follow him thereby lose some opportunities to score for the opposing team.

To illustrate an aspect of this issue, consider a famous article that tried to refute the claim that one sometimes observes a "hot hand" in basketball. Basketball players, coaches, and fans all tend to believe that a player can be "on a roll" where he can do no wrong, scoring from the most difficult positions. The authors argued that the belief was a myth, due to a

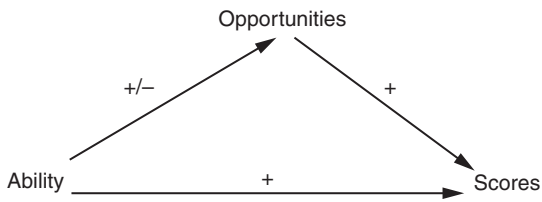



Figure 10.4

misunderstanding of the operation of chance. Statistical analyses showed that the outcome of shots was largely independent of the outcome of the previous attempt. The finding might, however, be compatible with the existence of a hot hand if the opposing team, correctly observing that the player is on a roll, neutralizes his higher ability to score by reducing his opportunities.[18]

In many contexts, higher ability obviously increases opportunities. A good applicant to college may be accepted at five institutions, whereas a less able student may be accepted at one or none. As in individual sports, the situation is competitive, but applicants have no opportunity to affect the opportunities of their rivals. Some societies and subcultures seem to punish excellence by ostracizing outstanding individuals or blocking all doors to them. I give some examples in Chapters 24 and 25.

The relation between desires and abilities can take several forms: a strong desire can reduce the ability, a weak ability can reduce the desire, and a strong ability can enhance the desire.

The first effect is illustrated by an example cited by Montaigne: an "excellent bowman was condemned to death, but offered a chance to live if he would agree to demonstrate some noteworthy proof of his skill. He refused to make an assay, fearing that the excessive strain on his will would make his hand go wrong and that instead of saving his life he would also lose the reputation that he had acquired as an archer." It has often been noted, in fact, that archery and rifle shooting require an almost Zen-like deceleration of the heart rate, to enable the performer to release the string or pull the trigger between the beats. If the stakes are high, this state may be hard to achieve.

The second effect is illustrated by a variety of the "sour-grapes" mechanism. As we have seen, the lack of opportunity to do X may stifle the desire to do X. The lack of ability to do X (given the opportunity) may have the same effect. We tend naturally to upgrade the importance and value of the activities at which we excel, and to downgrade those in which we do poorly. Since for each person in any pair of individuals there is probably one activity at which he or she is better than the other, this transmutation or "life lie" (Chapter 9) may counteract envy. In the long run, though, our amour-propre may be undermined by the lack of appreciation by others of our trivial achievements.

The third effect is illustrated by artists who suffer from excessive ability: they prefer to do what only they can do, at the expense of their art. The jazz singer Sarah Vaughan is one example of the dangers of virtuosity, and there are many others. In much of ancient philosophy – Aristotle and Seneca are

[18] The authors anticipate this objection by asserting that their claim also holds for free throws, which the opposing team is unable to block. Their answer does not affect my conceptual point. In any case, I agree with critics who observed that the uncontroversial existence of "cold hands" (perhaps literally due to a cold) implies the existence of hot hands as well.

examples – it was a given that human beings should cultivate and deploy only the faculties that only human beings possess. Marx took over this idea, when he asserted that self-realization through creative work (see earlier quotation) is the essence – "the species-being" – of humanity.[19] Although these are normative rather than explanatory claims, they can explain behavior if agents take them to heart and act upon them.

Abilities can also have opportunities as their *object*. Having an opportunity is useless if you do not know that you have it. Agents may be more or less able to identify the opportunities. A doctor who has not followed the medical literature since graduating from medical school will be less able to identify opportunities for intervention than one who has kept up to date. New drugs, for instance, may have appeared on the market. Also, having an opportunity is useless if you do not know the consequences of choosing it. Agents may be more or less able to estimate these consequences. The same example illustrates this point, if, for instance, the first doctor is unaware of the side effects of a given drug. These cognitive abilities differ from what we usually think of as physical skills, such as the ability to perform complicated surgery or the ability to score on a basketball shot.

### Bibliographical note

The constraints on voting in Massachusetts are cited from C. Williamson, *American Suffrage* (Princeton University Press, 1960), p. 273. The idea of "irresistible desires" is effectively demolished by G. Watson, "Disordered appetites: addiction, compulsion, and dependence," in J. Elster (ed.), *Addiction: Entries and Exits* (New York: Russell Sage, 1999). The claim that all individuals have the same preferences and differ only in the opportunities they face is notably associated with G. Stigler and G. Becker, "De gustibus non est disputandum," *American Economic Review* 67 (1977), 76–90. I take the idea of strategic misrepresentation of opportunities by Roman army commanders from S. Phang, *Roman Military Service* (Cambridge University Press, 2008), p. 68. The idea of adaptive preferences and notably of overadaptation to constraints is due to P. Veyne, *Le pain et le cirque* (Paris: Seuil, 1976) (partial translation in *Bread and Circuses* [New York: Penguin, 1982]). The decline in the (overall) suicide rate that followed the switch from coal gas to natural gas in Britain is documented in N. Kreiman, "The coal gas story," *British Journal of Social and Preventive Medicine* 30 (1976), 86–93. The extent to which reduced access to one means of suicide induces greater use of other means is

[19] The assertion is somewhat paradoxical, since it seems to devalue the activity of reading the books, looking at the paintings, or listening to the musical pieces that are the vehicles of self-realization. At most, the audience would consist of fellow artists.

discussed in C. Cantor and P. Baume, "Access to methods of suicide: what impact?" *Australian and New Zealand Journal of Psychiatry* 32 (1998), 8–14. The two examples of the effect of the packing of paracetamol are taken from D. Gunnell *et al.* (1997), "Use of paracetamol for suicide and non-fatal poisoning in the UK and France: are restrictions on availability justified?" *Journal of Epidemiology and Community Health* 51 (1997), 175–79, and J. Turvill, A. Burroughs, and K. Moore, "Change in occurrence of paracetamol overdose in UK after introduction of blister packs," *The Lancet* 355 (2000), 2048–9. Madison's use of the opportunity–desire distinction is analyzed by M. White, *Philosophy*, The Federalist, *and the Constitution* (New York: Oxford University Press, 1987). I discuss Tocqueville's use of the opportunity–ability–desire distinction in Chapter 5 of *Alexis de Tocqueville: The First Social Scientist* (Cambridge University Press, 2009). The observations on slavery in the Old South are from A. Taylor, *The Internal Enemy* (New York: Norton, 2013), pp. 100, 402. Studies of "salutary constraints" in the arts include Chapter 3 of my *Ulysses Unbound* (Cambridge University Press, 2000) and T. Osborne, "Rationality, choice, and modernism," *Rationality and Society* 23 (2011), 175–97. The study of the "hot hand" in basketball is T. Gilovich, R. Vallone, and A. Tversky, "The hot hand in basketball: on the misperception of random sequences," *Cognitive Psychology* 17 (1985), 295–314, critically discussed in K. Korb and M. Stillwell, "The story of the hot hand: powerful myth or powerless critique?" International Conference on Cognitive Science (2003). For what Paul Veyne calls the "zoological snobbism" of Aristotle and Seneca, see his annotations to *Sénèque* (Paris: Robert Laffont, 1993), pp. 814, 925, 1178. I discuss the dangers of virtuosity in jazz in *Ulysses Unbound*, pp. 247–52.

# 11 Reinforcement and selection

The emphasis on choice as the basic building block of explanation in the social sciences goes together with an emphasis on the *intended* consequences of action.[1] In this chapter, I discuss explanations of actions in terms of their *objective* consequences. This might seem like an unpromising idea. All explanation is causal explanation. We explain an event by citing its cause. Causes precede their effects in time. It follows that we cannot explain an event, e.g. an action, by its consequences.

If, however, the explanandum is a *pattern* of recurrent behavior, the consequences of that behavior on one occasion can enter into the causes that make its occurrence on a later occasion more likely. There are two main ways in which this can happen: by *reinforcement* and by *selection*. I shall focus on the second, which is the more important for my purposes, but begin with some words about the first.

## Reinforcement

If the consequences of given behavior are pleasant or rewarding, we tend to engage in it more often; if they are unpleasant or punishing it will occur less often. The underlying mechanism could simply be conscious rational choice, if we *notice* the pleasant or unpleasant consequences and *decide* to act in the future so as to repeat or avoid repeating the experience.[2] Often, however, the reinforcement can happen without intentional choice. When infants learn to cry because the parents reward them by picking them up when they do, there is no reason to think that they first consciously note the benefits from crying and then later cry at will to get them. When older children throw a tantrum to get their way, parents can usually tell that it is not a genuine one.

[1] In Chapter 4, I distinguished between consequentialist and non-consequentialist actions. For present purposes, the action itself may be included among its consequences. This is a purely terminological issue.

[2] Recall, however, that we are not always very good at noticing which of two experiences was the more painful (Chapter 6).

Reinforcement learning has been extensively studied in laboratory experiments on animals. One typically offers the animal the opportunity to press a lever, or one among several levers, and rewards the presses either as a function of the number of lever presses since the last reward or as a function of the time passed since the last reward. In either case, the function can be deterministic or probabilistic. In *fixed-ratio* schedules, the animal receives a reward after it has pressed a lever a fixed number of times, whereas in *variable-ratio* schedules the number of presses needed to produce a reward varies randomly. In either case, each press produces a "reward point" that is added to previous points. In *fixed-interval* schedules a press will produce a reward a given time after the last reward was offered, whereas in *variable-interval* schedules the time before a new reward is made available varies randomly. In either case, the timing of the reward is independent of the number of presses. Each schedule of reinforcement produces, after some learning time, a specific and stable pattern of behavior, which, moreover, will be extinguished in a specific pattern once the reinforcer (the reward) is removed. For instance, responses that are learned by rewarding every lever press (a special case of a fixed-ratio schedule known as continuous reinforcement) are extinguished more quickly than those learned on a random variable-ratio schedule. Intuition might suggest the opposite, since continuous reinforcement would seem to produce a stronger habit, but, as sometimes happens, intuition is wrong.

The relevance of these findings outside the laboratory depends on the purpose. If the aim is to *shape* action, for example, in a classroom situation, in a gambling casino, or in the workplace, a designer may (more or less freely) impose a reward schedule to generate desired behavior. For instance, variable-interval schedules are often used to shape behavior, as when a teacher announces a policy of random quizzes. On the variable-ratio schedule that operates in many gambles, it is easier to establish the behavior if the first reward occurs early on.[3] As casino and race track managers lack the technology for sucking in novices by offering them big wins, they have to rely on the fact that by the laws of chance some gamblers will have beginner's luck.[4] Con-man operations, however, often rely on the deliberate inducement of early wins by the mark. In the classroom and the casino, the reward schedules operate "behind the back" of the students or gamblers, in the sense that they do not shape the behavior by explicit incentives but rather, as in the case of the crying infant, by an unconscious process. By contrast, when managers pay employees once they achieve a set target (a fixed-ratio schedule) or on a monthly basis (a fixed-interval schedule), they are simply setting up an

[3] It is also easier to establish when the gambling technology allows for the possibility of near-wins. Although each of the near-wins is less reinforcing than an actual win, there are more of them.

[4] Their good luck, in this case, is their bad luck, and the casino's good luck.

incentive system (Chapter 25). Since the behavior of the employees can be adequately explained by the *expected* reward, there is no need to appeal to *actual* reward.

If the aim is to *explain* behavioral patterns by their actual consequences, the reward schedules are relevant only if they occur naturally and, moreover, are so opaque that they do not create explicit incentives. This does not often seem to happen with the two fixed schedules. In everyday life, the sheer number of responses is rarely decisive for reward. It is not the number of friendly smiles I give my friends that shapes their behavior toward me, but the consistency and the appropriateness of my smiling. In natural settings, rewards that arrive every so often, such as my paycheck, are rare. The two variable schedules are more important. A person who plays "hot and cold" (a variable-ratio schedule) with a member of the opposite (or the same) sex may induce a stronger attraction than someone who invariably displays friendly behavior. A variable-interval schedule arises when you try to reach someone on the phone and the line is busy. You know that sooner or later you will get through by redialing, but you do not know when. This situation induces a pattern of steady redialing that would not be the unique prediction of rational-choice theory. That theory could predict any number of patterns, depending on the caller's beliefs about how long the conversation of the other person is likely to last. It seems unlikely, however, that people have stable beliefs about such matters.

The response pattern generated by reinforcement is not, in general, the one that would be produced by conscious, rational choice. Suppose for instance that an animal has the choice between pressing either of two levers, one that rewards on a variable-ratio schedule and one that rewards on a variable-interval schedule. The rational pattern, which will maximize overall reward, is to press the variable-ratio lever most of time, to accumulate reward points, while visiting the variable-interval lever from time to time to see whether a new reward has become available. This is not, however, the pattern produced by reinforcement learning. Instead, the animals press the variable-interval lever much more often than is optimal. In doing so, they equalize the *average* rewards to pressing the one or the other lever rather than, as rationality would dictate, equalizing the *marginal* rewards. For other schedule combinations, reinforcement learning sometimes mimics rational choice, but not in any consistent manner. If there is any non-intentional mechanism capable of reliably simulating rationality, we shall have to look for it elsewhere.

### Natural selection

In most of this book I discuss how we can explain behavior by assuming that *agents adapt to their environment*, in a more or less rational manner. In a radically different perspective, we may try to explain behavior by assuming

that *agents are selected by the environment*. Although selection can be the work of an intentional agent, as when domestic dogs are bred to be docile or laboratory rats to become more intelligent, many selection mechanisms rest on causal processes that involve no intentional agent. In particular, *differential survival* of organisms based on their behavioral patterns may lead to optimal behavior (optimal for reproduction) in the population even in the absence of any optimizing choices or intentions. Suppose that 10 percent of the organisms in a population of 100 organisms forage so efficiently that they leave on average 10 offspring that survive to adulthood, whereas the remaining 90 percent leave only 5. If the behavior of the parents is (by whatever mechanism) transmitted to the offspring, the next generation of adult organisms will include a fraction of 100/550 = 2/11 ~ 18 percent that displays the more efficient behavior. Over the course of a few more generations, virtually all organisms will display it. If we ask *why* it is universally displayed, the answer is that it has better consequences. This mechanism works across generations. Unlike reinforcement learning, it does not modify the behavior of any given individual, only the typical behavior of successive generations of individuals.

This story does not say *why* some organisms forage more efficiently than others. The generally accepted answer is that changes in the structure or (as in this example) the behavior of organisms are due to mutations in the genetic material that occur when the gene is copied from one generation to the next. The mutations have four important properties. First, they arise more or less at random. Although mutagens can increase the frequency of mutations, just as a poor-sighted typesetter will make more mistakes, there is no mechanism that would favor beneficial or useful mutations, any more than the mistakes of an ignorant typesetter could systematically improve the text he is setting. Second, most mutations are in fact deleterious, that is, the protein for which the mutated gene codes reduces reproductive fitness. Most mistakes by a typesetter also create confusion. Third, mutations are typically small, involving a substitution of one "letter" (one of four nucleotide molecules) in the genetic code for another. A mistake by a typesetter, too, often involves the replacement of one letter by another, if for instance she substitutes "366 days" for "365 days." Fourth, some mutations *do* improve fitness. In a leap-year, the typesetter's mistake might yield an improvement. Natural selection is based on happy accidents of this kind.

Because mutations are small, natural selection is constrained to take one step at a time. Each step has to be viable, since otherwise the organism in which the mutation occurred would not leave any descendants. Natural selection (in this classical picture) is constrained to small incremental improvements. The organism climbs along a fitness gradient until it reaches a *local maximum*, defined as a state in which all further one-step changes would reduce fitness.

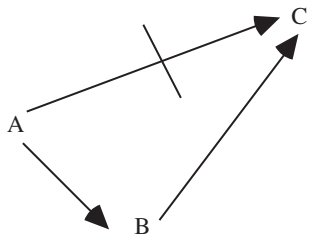



Figure 11.1

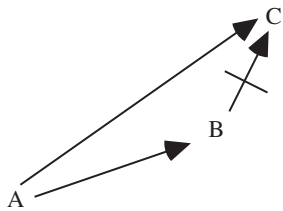



Figure 11.2

Although there may be higher peaks in "the adaptive landscape," these will not be attainable by one-step changes.

This process differs from intentional choice in three respects. Recall from Chapter 6 that by virtue of their intentionality, human beings are capable of (1) using indirect strategies, (2) waiting, and (3) aiming ahead of a moving target.

Concerning (1), natural selection cannot *reculer pour mieux sauter,* since an organism that took one step backward (a deleterious mutation) would not leave offspring in which a favorable mutation (two steps forward) could occur. Figure 11.1 illustrates this case. A direct move from A to C is impossible, because it would require the simultaneous mutation of two nucleotides. While a one-step move from A to B is possible, the further one-step move from B to C is blocked by the fact that an organism in state B would not leave any descendants to take the second step.

Concerning (2), natural selection cannot *wait*, that is, refuse a favorable mutation from A to B if it would preclude a move to a global maximum C that can be reached from A, but not from B. Figure 11.2 illustrates this case.

Concerning (3), populations are adapting to an environment that is constantly changing. If the changes are regular, for instance, seasonal or diurnal, they adapt to the changes. If a one-shot event occurs, such as a sudden climate change, behavior that was at a local fitness maximum prior to the change may become suboptimal, so that mutations that previously would have been

deleterious are favorable. If the change is protracted, as when the climate cools or warms over a long period of time, this process may never reach a new local maximum. The population will track the changes in the environment with an efficacy that depends on the relative speed of the two processes. The amazingly fine-tuned adaptations observed in animals and plants suggest that animals adjust to the environment much faster than the latter itself changes. Yet the organisms will always lag somewhat behind, since they cannot *anticipate* changes in the environment. By contrast, human beings may become aware of future changes such as global warming and take precautions against them before they actually occur or, if they result from human behavior, prevent them from occurring.

The environment of a population is made up, among other things, of populations from other species to which it may stand in a prey-predator relation. As prey, it may evolve better evasive strategies; as predator, better hunting strategies. Just as the individual fox and hare are chasing each other across the fields, so do the species Fox and Hare chase each other over generations. But whereas the logic of natural selection precludes the Fox from anticipating where the Hare is going to be a few millennia hence, some predators are able to intersect the flight path of the prey (Figure 6.1). Similarly, the locally maximizing process of natural selection has produced the capacity for global maximization found in human beings.

This classical picture of natural selection is simplified in a number of ways. First, some behavioral patterns are transmitted over generations by culture, not by genes. Although the theory of "memes" as a cultural analogue of genes is too poorly specified to be useful, there exist persuasive models that explain, for instance, why women in some countries have fewer children than the biological maximum that would be favored by selection operating on genes alone. In contemporary Italy, the average number of children per woman is about 1.4. While the biological maximum is hard to define, it is certainly much higher. Second, large mutations do occur, and some of them may be responsible for developments that would not have been possible through small point mutations. Also, inferior forms are not eliminated instantly by competition. In Figure 11.1, the mutation to B does not necessarily produce an organism that is unviable in the strict sense of being unable to survive or to reproduce. Some organisms in state B might survive to produce organisms in state C. In Figure 11.2, some organisms in state A might survive the competition from the more efficient organisms in state B long enough for a mutation to state C to occur. Whether the global maximum occurs is a matter of the relative speed of two processes: the extinction of inferior varieties and the rate at which favorable mutations occur. There is no mechanism, however, that could mimic, *in a systematic way*, the capacity of intentional beings to preempt, to wait, or to use indirect strategies.

### The units of selection

Natural selection is not only opportunistic and myopic, but also, with two exceptions I shall discuss shortly, fiercely *individualistic*. It does not favor the species or the population, but the individual organism. If a property arising from a mutation increases the relative fitness of the organism, it will be fixed in the population even if it also causes a decrease in absolute fitness. Imagine a population of fish, exposed to predators, initially swimming in scattered schools. If a mutation causes the fish in which it occurs to move to the center of the school, it will be less vulnerable to predation and as a result will tend to leave more offspring. As this behavior spreads in the population, the school will become more compact and thus an easier target for predators. At each step in the process it is better to seek the middle than to be at the outskirts of the school. Yet in terms of absolute fitness the outcome is worse for all than the initial situation, and in terms of relative fitness it is unchanged. Similarly, runaway sexual selection is a plausible explanation of the large and dysfunctional antlers found in some species of deer.

One exception to individualism is *kin selection* (a form of "subindividualism"), in which the gene rather than the individual organism is the unit of selection. The choice of unit does not matter when the effect of a gene simultaneously and in the same proportion increases the presence of the gene in the population and the number of offspring of the organism displaying that behavior. This is the case, for instance, in the evolution of more efficient foraging. But in some cases the gene can benefit even if the organism in which it triggers the behavior does not, namely, when an organism "sacrifices" itself for the sake of close relatives who are likely to have the same gene. When an animal observes a predator and emits an alarm signal, its chances of survival often go down while those of close relatives in the vicinity may go up. Since these relatives or some of them will also have the "warning gene," their higher survival chances may cause the gene to spread in the population if they more than offset the lower chances of the animal that emitted the signal.

Another exception is *group selection* (a form of "supraindividualism"). Consider two populations of fish, one in which the center-seeking mutation has occurred and one in which it has not. Over time, the former will leave fewer offspring than the latter and might ultimately be crowded out. Selection would seem to operate at the level of the group, not of the individual. Yet if the two populations coexist, the second is vulnerable to invaders from the first. Whether the center-seeking behavior is caused by mutation or by in-migration, the outcome is the same, namely, the crowding out of those who do not behave in that way. Similarly, if organisms in a population have a gene that prevents them from overgrazing, thus avoiding "the tragedy of the commons," they might be out-reproduced by less inhibited organisms that lack the gene. For

this reason, group selection has not been seen as a plausible mechanism for generating cooperation or self-restraint. In the light of the theory of altruistic punishment set out in Chapter 5, however, this objection can be met. If the organisms in a population have a gene that make them punish non-cooperators, the latter will not gain any reproductive advantage from their free riding.[5]

Kin and group selection provide two mechanisms for the emergence of cooperative behavior, the former based on shared genes and the latter on altruistic punishment. A third mechanism is that of *reciprocal altruism* or "tit-for-tat" in repeated interactions, such as "I scratch your back; you scratch mine" (among some animals quite literally) or "I offer you food when I have a surplus; you offer it to me when you have." The other side of the coin is punishment, or at least abstention from cooperation, when the other party fails to reciprocate. For this mechanism to operate the individuals must interact often enough to make self-restraint worthwhile, remember what others did on earlier occasions, and recognize them when they meet again.

### Natural selection and human behavior

Natural selection has obviously shaped the physical structure of human beings, which offers them opportunities for action as well as constraints on action. Those who try to explain human behavior in terms of natural selection sometimes make stronger claims. They want to explain the behavioral patterns themselves, not merely the structures that make them possible.

The most plausible mechanism is that evolution has produced *emotions* with their characteristic action tendencies. Since a male can never be sure whether he is the father of his offspring whereas the female is not in doubt that she is the mother, we would expect natural selection to produce a stronger tendency to feel sexual jealousy in men than in women. This is confirmed by many homicide statistics. Thus among 1,060 spousal homicides in Canada between 1974 and 1983, 812 were committed by men and 248 by women, but among those motivated by jealousy 195 were by men and only 19 by women. The theory of natural selection also predicts that parents would be more emotionally committed to their biological children, who are carriers of their genes, than to stepchildren. This prediction, too, is confirmed by data. Thus an American child living with one or more substitute parents in 1976 was about 100 times as likely to be abused as a child living with two natural parents. Natural selection may also favor *lack* of emotion. The dangers of inbreeding are kept in check, in

[5] One might ask, however, whether a population of cooperating punishers might not risk being invaded by cooperating non-punishers, since punishing is costly for the punisher. The cooperating non-punishers might in turn be invaded by non-cooperators, and a new cycle might start. To my knowledge, this conundrum is not fully resolved.

humans and in other primate species, by lack of sexual attraction among young who grow up together, whether or not they are related to each other.[6]

Natural selection, operating on groups rather than on individuals, may also have favored emotions of anger and indignation toward those who violate norms of cooperation, motivating punishment even at some cost to the punisher. A more puzzling question is whether and why selection might have favored the emotion of *contempt*, which is directed toward those who violate social rather than moral norms. Since many social norms are arbitrary and even dysfunctional, it is difficult to see how they could be sustained by group selection. Given a tendency for others to ostracize those who violate social norms, reproductive fitness would be better served by respecting the norms if the cost of ostracism exceeds the benefit derived from the norm violation. The puzzle is why this tendency would arise in the first place. Why, for instance, would people disapprove of adultery? Social norms against adultery involve third-party reactions that differ from the second-party reaction of sexual jealousy. Although A might benefit from C's disapproval of B's advances to A's spouse, that benefit does not create a selection pressure on C to behave in this way. Whereas group selection might favor genes that induce "third-party punishment" of free riders, the benefit to the group of third-party punishment of adultery is less obvious. Although the tendency for norms against female adultery to be stronger than those against male adultery suggests an evolutionary explanation, it is hard to see what the mechanism would be.

Other claims are more speculative, such as the idea that *self-deception* in humans evolved because of its evolutionary benefits. The argument goes as follows. It is often useful to deceive others. However, deliberate or hypocritical deception is hard to carry off. Therefore, self-deception evolved to enable people to deceive others successfully. The weakness of the argument is that if self-deception causes one to hold false beliefs, these might have disastrous consequences if used as a premise for behavior. Nobody has made a convincing argument that the *net* effect of these opposing effects tends to be positive, as it would have to be for self-deception to enhance evolutionary fitness.

Even more speculative is the claim that unipolar *depression* may have evolved as a bargaining tool, somewhat similar to a labor strike. For instance, an alleged function of postpartum depression is to induce others to share in raising the child, just as workers go on strike to make employers share the profits. Suicides induced by depression are, on this view, the cost of making a credible threat of suicide. They are, as it were, suicide attempts that failed to fail. Insomnia is explained as an allocation of cognitive resources to solve the

[6] The incest taboo may, therefore, address a temptation that exists more rarely than has been thought. Freud, by contrast, thought the incest taboo had arisen to counteract an unconscious desire to have sex with close relatives.

crisis to which the depression is a response, whereas hypersomnia (sleeping more than normal) is explained as a way of reducing productivity and thus enhancing the bargaining efficacy of the depression. The argument, while consistent with some known facts about depression, ignores a host of others, such as that depression as well as suicide run in families, that divorced individuals (with no bargaining partner) are more depression prone than the married or never married, and that stressful life events are neither necessary nor sufficient for depression.

Explaining depression as a bargaining tool is another example of a pervasive search for a *meaning* or *function* of all apparently pointless or dysfunctional behaviors (see Chapter 9). Up to a point, the search for meaning is a good research strategy; beyond this point, it becomes contrived and, as in some of the examples cited, ultimately absurd. There are so many ways in which harmful traits may be preserved in a population that one cannot take for granted that frequently occurring behavior confers reproductive fitness on the agent.[7] Natural selection has certainly favored the propensity to feel physical pain, and there is no a priori reason why it could not favor the tendency to experience mental pain. But to establish the function of depression it is not enough to offer a just-so story that accounts for some of the known features of the illness. Crucially, the hypothesis must also explain facts over and above those it was constructed to explain (Chapter 1), and preferably "novel facts" that were unknown until predicted by the hypothesis.

## Variation and selection

Up to this point I have assumed the standard biological case of *random* variation and *blind* deterministic selection. Selection models can take other forms, however, involving intentional variation, intentional selection, or both (see Figure 11.3].

### *Intentional variation, intentional selection*

In *The Origin of Species*, Darwin wrote that "nature gives successive variations; man adds them up in certain directions useful to him." But it is not merely a case of "Nature proposes; man disposes," since, as he also observed, the input can be modified by human behavior:

[7] A given gene may code for several behaviors (pleiotropy) and be maintained even if one of them by itself is suboptimal. Suboptimal features may also be maintained by a variety of other genetic mechanisms related to the fact that sexually reproducing organisms have two different variants (alleles) of each gene.

| | Intentional source of variation | Nonintentional source of variation |
|---|---|---|
| Intentional selection | Artificial selection in plant and animal husbandry | Gradual improvement of boats<br>Eugenics<br>Selective abortion and infanticide |
| Nonintentional selection | Selection of firms by market competition | Natural selection |

Figure 11.3

A high degree of variability is obviously favourable, as freely giving the materials for selection to work on; not that mere individual differences are not amply sufficient, with extreme care, to allow of the accumulation of a large amount of modification in almost any desired direction. But as variations manifestly useful or pleasing to man appear only occasionally, the chance of their appearance will be much increased by a large number of individuals being kept; and hence this comes to be of the highest importance to success. On this principle Marshall has remarked, with respect to the sheep of parts of Yorkshire, that "as they generally belong to poor people, and are mostly in small lots, they never can be improved." On the other hand, nurserymen, from raising large stocks of the same plants, are generally far more successful than amateurs in getting new and valuable varieties.

Today, we can add that artificial selection can also be enhanced by inducing mutations. In addition, the maintenance of "genetic libraries" can prevent the reduction of genetic variation that is otherwise the inevitable result of selection for particular traits.

With regard to the selection process itself, Darwin distinguished between two *levels of intentionality*:

At the present time, eminent breeders try by methodical selection, with a distinct object in view, to make a new strain or sub-breed, superior to anything existing in the country. But, for our purpose, a kind of Selection, which may be called Unconscious, and which results from every one trying to possess and breed from the best individual animals, is more important. Thus, a man who intends keeping pointers naturally tries to get as good dogs as he can, and afterwards breeds from his own best dogs, but he has no wish or expectation of permanently altering the breed.

*Non-intentional variation, intentional selection*

There are many cases in which a new organism or a new form arises by accident and is then either accepted of rejected on the basis of intentional choice. Whereas natural selection tends to produce a roughly equal number of male and female organisms, gender-biased infanticide and more recently gender-biased abortion can create a serious sex imbalance in the population. In India and China alone, around 80 million women are "missing" for this reason. Eugenic policies have been widely used to prevent the mentally ill and mentally retarded from reproducing. In Nazi Germany, around three hundred thousand to four hundred thousand individuals were forcibly sterilized on these grounds. As prenatal screening techniques improve, selective abortion may become an important determinant of the makeup of human populations. If further advances make it possible to determine the sex of the child at conception, selection will have been replaced by intentional choice.

Random variation combined with intentional selection may also shape the development of artifacts. When the Norwegian minister and sociologist Eilert Sundt visited England in 1862, he learned about Darwin's theory of natural selection (published in 1859) and set about applying a variant of it to boat construction:

> A boat constructor may be very skilled, and yet he will never get two boats exactly alike, even if he exerts himself to this end. The variations arising in this way may be called *accidental*. But even a very small variation usually is noticeable during the navigation, and it is then *not accidental* that the seamen come to *notice* that boat that has become improved or more convenient for their purpose, and that they should recommend this to be *chosen* as the one to *imitate* ... One may believe that each of these boats is perfect in its way, since it has reached perfection by one-sided development in one particular direction. Each kind of improvement has progressed to the point where further developments would entail defects that would more than offset the advantage ... And I conceive of the process in the following way: when the idea of new and improved forms had first been aroused, then *a long series of prudent experiments*, each involving extremely small changes, could lead to the happy result that from the boat constructor's shed there emerged a boat whose like all would desire.

In this text, Sundt improved on Darwin in a crucial respect.[8] Whereas Darwin confessed to ignorance about the origin of variation, Sundt hit on the idea of

[8] The improvement was possible, of course, only because he addressed a different problem, since in 1862 nobody had the conceptual wherewithal to imagine that the source of variation in *organisms* could be random replication mistakes. This leap became possible only after Mendel had shown the discrete nature of the units of inheritance (genes) and Watson and Crick demonstrated that replication was involved in the process of inheritance. I wonder what Darwin would have answered had Sundt asked him whether the source of biological variation might not be imperfection in the reproductive machinery.

locating its source in *errors of replication*, similar to typographical errors and to (what we know now to be) mutations in the genetic material. The imperfection of the boat builder – his inability to make perfect copies – is a condition for the perfection of the end result. Sundt carefully notes that the outcome of the process is a local maximum, from which no further improvements can occur by incremental changes. In the very last sentence, he also suggests that the process may turn into artificial selection, when people engage in deliberate experiments rather than letting variations arise by chance. As did Darwin, he suggested that intelligence or intentionality may occur at two levels: first when people *notice* than one model is more seaworthy than a previous one, and then when they *understand* that improvements could be accelerated if chance variation were replaced by systematic experiments.[9]

### *Intentional variation, non-intentional selection*

The working of economic markets has some features in common with natural selection. The analogy has two versions, one relatively close to natural selection and one more remote. They share the premise that given the multiple limitations of human rationality, firms or managers are *inefficient* in the sense that they are unable to calculate the production and marketing decisions that will maximize their profit. Nevertheless the market mechanism will weed out inefficient firms, so that at any given time mainly efficient firms will be observed. Everything happens "as if" managers were efficient.

In the first and simplest version, all firms are constantly trying to increase their profits by processes of imitation and innovation. Although imitation by itself does not generate new inputs for selection to operate on, *imperfect* imitation may, as noted, have this result. Innovation is also, by definition, a source of new inputs. When – through sheer luck – innovation or imperfect imitation enables a firm to produce at lower cost, it can undersell its rivals and drive them out of business unless they, too, adopt the more efficient ways. By the mechanisms of bankruptcy, takeover, and imitation, these efficient techniques will spread in the population of firms. If we assume that both imitation and innovation occur predominantly in small steps and that competition takes place in an otherwise constant environment, it will bring about a local maximum of equilibrium profits.

The second version denies that firms are always trying to maximize profits. Instead, they use *routines* or rules of thumb that are maintained as long as profits are at a "satisfactory" level. In a neologism, they "satisfice" rather than maximize. What this means may depend on many factors, but we may assume

[9] In this way, one could also prevent the unfortunate situation that would arise if boat builders became so good that they never made mistakes.

for simplicity that a firm whose profits are consistently below the satisfactory level will either go bankrupt or face the threat of a hostile takeover. The simplest routine is to do everything as before as long as profits are at a "satisfactory" level. More complicated routines could include setting prices by a constant markup on costs or investing a certain percentage of profits in new production. The idea of satisficing is reflected in such sayings as "Never change a winning team" or "If it ain't broke, don't fix it." In one perspective, satisficing could even be optimal. In a phrase I quoted earlier, "The greatest of all monopoly profits is a quiet life."

Suppose now that profits fall below the satisfactory level. A firm that has been doing the same thing year in year out may be the victim of an organizational analogue of rust or sclerosis. External shocks such as a rise in oil prices or a change in an important exchange rate may increase costs or reduce revenue. Consumer demand may change; rivals may come up with better methods or new products; or workers might impose a costly strike on the firm. Whatever the cause (which may even be unknown to the firm), unsatisfactory profits will induce a search for new routines by some combination of innovation and imitation. Either procedure is likely to be predominantly local, in the sense of being limited to alternatives close to the existing routines. Large changes of any kind may be too costly for a firm that is in financial trouble (Chapter 10), and non-incremental innovations are also conceptually more demanding.

The process of imitation is obviously biased toward the behavior of successful rivals. Whether innovation is random or directed depends on the perceived causes of the crisis that triggered it. If the fall in profits below the acceptable level resulted from a rise in oil prices, the firm may bias its search in the direction of methods that will economize on oil.[10] If it resulted from a change in the exchange rate between the dollar and the euro, the firm is more likely to search randomly. In all cases, however, there is a strong intentional component in the firm's behavior. The decision to change the current routines is intentional, as is the decision about how much to invest in innovation or in imitation. The choice of models to imitate is deliberate, and as just noted, the firm may intentionally bias the search for new routines in a particular direction.

The new routines that result from this process are then exposed to the blind forces of market competition. If they enable the firm to attain a satisfactory level of profit, it will switch off the search until a new crisis arises. If they do not, the firm may try again or have to declare bankruptcy. Sooner or later, non-

[10] Unless further increases in the price of oil are to be expected, rationality does not require the firm to look for *oil-saving* innovations, since no one can know what the set of feasible innovations looks like. Yet the increase in the price of oil will tend to make these innovations more salient.

satisficing firms are eliminated. In itself, this process does not tend to produce profit-maximizing firms. To see how that could happen, we need to bring *competition* more explicitly into the picture. If we assume that as one of their routines firms invest a fixed percentage of profits in new production, those that by sheer luck have hit upon a better routine than their competitors will expand so that over time their routines become more heavily represented in the population of firms.[11]

### Selection and as-if rationality

The usefulness of these models depends on a simple empirical question: what is the rate at which inefficient firms are eliminated compared to the rate of change of the environment? In the previous chapter I raised the same question with regard to natural selection and argued that the highly fine-tuned adaptation of organisms to their environment suggests that the latter must have changed relatively slowly. In the case of the economic environment, we can make a more direct assessment. In the modern world, firms are exposed to unprecedented rates of change. If they were reduced to incremental tracking of the environment, firms would be chronically unfit. Successful firms are more likely to be those that are capable of *anticipating* change, by aiming ahead of the target. This strategy, too, will fail much of the time, but at least not all the time. Moreover, because of their political clout large corporations may also be able to *shape* the environment in which they operate. In an earlier age of cutthroat capitalism among small firms, selection mechanisms of the kind I have described may or may not have been important – we do not know. Today, they are unlikely to explain much of what we observe.

There is also a more general issue at stake. When attacked for the lack of realism of their assumptions, rational-choice theorists routinely assert that they only claim to explain behavior on the assumption that people act "as if" they maximize utility (or profit, or any other aim). Milton Friedman offered two seductive and influential analogies to persuade his readers of the reality of maximizing behavior that does not rely on maximizing calculations. First, "leaves [on a tree] are positioned as if each leaf deliberately sought to maximize the amount of sunlight it receives, given the position of its neighbors, as if it knew the physical laws determining the amount of sunlight that would be received in various positions and could move rapidly or instantaneously from any one position to any other desired and unoccupied position." Second, "excellent predictions would be yielded by the hypothesis that the

[11] In this version they will not deliberately try to drive their rivals out of the market, for instance, by using the high profits to sell below cost until the others give up, since they have no concern for more-than-satisfactory profits.

[expert] billiard player made his shots as if he knew the complicated mathematical formulas that would give the optimum directions of travel, could estimate accurately by eye the angles, etc., describing the location of the balls, could make lightning calculations from the formulas, and could then make the balls travel in the direction indicated by the formulas."

While seductive, the analogies are unpersuasive. The leaves simulate maximization because natural selection eliminated trees that did not. To assume that a similar mechanism exists for economic behavior is to beg the question. Expert billiard players are experts because ten thousand hours of practicing enable them somehow (we do not know how) to make the right shots on an intuitive basis. Chess grandmasters can instantly recognize about 50,000 constellations on the board. These are, of course, tightly constrained situations. To extrapolate the argument to business decisions in a fluid and opaque environment is unwarranted.

The most general way of stating my objection is perhaps that even if it could be shown that market competition does improve efficiency through elimination of inefficient firms, there is a vast step from "improving efficiency" to the ultrasophisticated as-if maximization imputed to firms in economic models.[12]

In the political sphere, electoral competition is supposed to ensure that the only politicians we observe are those who are elected or reelected; hence one can assume that all politicians act "as if "they are concerned only with their election prospects. The leap from a concern with election to an *exclusive* concern is not justified, however. A methodologically unprejudiced look at politics suggests that there are three kinds of political actors: opportunists (who care only about getting elected and reelected), reformers (who care about their policies being implemented), and activists or militants (who care more about "making a statement").[13] The view of politics as based on the interaction among these three groups in each party – and among different parties – is clearly more realistic than the "ice cream stall" model of politics (Chapter 18) according to which vote-maximizing parties would all converge to the center.

[12] As a small wrinkle to the argument, the economics of team sports offers a possible objection to the idea that profit maximization is brought about by selection. If profit-maximizing baseball or football teams used their profits to buy up all the best players in their league, their superiority would become so overwhelming that the games would lose much of their uncertainty, and hence of their fun, and hence of their profit-generating ability.

[13] The three groups can be more formally distinguished as follows. Opportunists prefer to propose policy A to policy B when the probability of winning at A is greater than the probability of winning at B, given that the opposition party is proposing some fixed C. Militants prefer to propose A to B when the average party member would derive higher utility at A than at B (independently of what C is). Reformists prefer to propose A to B, given that the opposition is proposing C, when the expected utility of the average party member is higher at A than at B. Thus opportunists are concerned only with probabilities, activists only with utilities, and reformers with both.

For a striking refutation of the claim that politicians are motivated only by reelection concerns, consider the line of French politicians originating in Jean Jaurès and passing through Léon Blum, Pierre Mendès-France, and Michel Rocard, all of whom were transparently motivated by a desire to promote the impartial values of social justice and economic efficiency. It has to be said, though, that in Rocard's case his distaste for electoral politics did detract from his political efficacy.

Outside the arenas of competition, "as-if" rationality has even less justification. Consumer choices, voting behavior, church attendance, choice of career, and most other behaviors one could name are not subject to selection mechanisms that mimic rationality. They are, to be sure, subject to *constraints* that can reduce the importance of choice in general and of rational choice in particular (Chapter 10). Constraints operate before the fact, to make certain choices unfeasible. Selection operates after the fact, to eliminate those who have made certain choices. Although both mechanisms contribute to the explanation of behavior, they cannot, jointly or singly, account for all of it. Choice remains the core concept in the social sciences.

### Bibliographical note

In "Selection by consequences," *Science* 213 (1981), 501–4, B. F. Skinner argued for the importance of *three* ways in which behavior can be explained by its consequences: by natural selection operating on individuals, by reinforcement, and (although he does not use that term) by group selection. A useful introduction to reinforcement theory is J. E. R. Staddon, *Adaptive Behavior and Learning* (Cambridge University Press, 1983). A study of how reinforcement theory can be used to shape (rather than explain) behavior is D. Lee and P. Belflore, "Enhancing classroom performance: a review of reinforcement schedules," *Journal of Behavioral Education* 7 (1997), 205–17. A classic exposition of the theory of natural selection, notable for the insistence on the individualistic nature of selection, is G. Williams, *Adaptation and Natural Selection* (Princeton University Press, 1966). For a discussion of gradient climbing and "the metaphor of fitness landscapes," see Chapter 2.4 of S. Gavrilets, *Fitness Landscapes and the Origin of Species* (Princeton University Press, 2004). An exposition emphasizing the gene as the unit of selection is R. Dawkins, *The Selfish Gene*, 2nd edn (Oxford University Press, 1990). An excellent introduction to animal signaling is S. A. Searchy and S. Nowicki, *The Evolution of Animal Communication* (Princeton University Press, 2005). For a discussion of how group selection might be made possible by altruistic punishment, see E. Fehr and U. Fischbacher, "Social norms and human cooperation," *Trends in Cognitive Sciences* 8 (2004), 185–90. A seminal study of "tit-for-tat" cooperation between unrelated animals is R. Axelrod and

W. Hamilton, "The evolution of cooperation," *Science* 211 (1981), 1390–6. The data on homicide statistics and child abuse are from M. Daly and M. Wilson, *Homicide* (New York: Aldine de Gruyter, 1988). For objections to their explanation, see Chapter 7 of D. Buller, *Adapting Minds* (Cambridge, MA: MIT Press, 2005). For two sides of the self-deception argument, see R. Trivers, *Social Evolution* (Menlo Park, CA: Benjamin-Cummings, 1985) (favoring an evolutionary explanation), and V. S. Ramachandran and S. Blakeslee, *Phantoms in the Brain* (New York: Quill, 1998) (opposing it). For two sides of the adaptive nature of depression, see E. H. Haggen, "The bargaining model of depression," in P. Hammerstein (ed.), *Genetic and Cultural Evolution of Cooperation* (Cambridge, MA: MIT Press, 2003) (favoring an evolutionary explanation), and P. Kramer, *Against Depression* (New York: Viking, 2005) (opposing it). The analysis of markets in terms of natural selection originates in A. Alchian, "Uncertainty, evolution, and economic theory," *Journal of Political Economy* 58 (1950), 211–21. Its most sophisticated version (which does *not* support "as-if" maximization) is R. Nelson and S. Winter, *An Evolutionary Theory of Economic Change* (Cambridge, MA: Harvard University Press, 1982). The theory of "satisficing" derives from H. Simon, "A behavioral theory of rational choice," *Quarterly Journal of Economics* 69 (1954), 99–118. The economics of team sports is the subject of D. Berri, M. Schmidt, and S. Brook, *The Wages of Wins* (Stanford University Press, 2006). The distinction among opportunists, reformers, and activists is taken from J. Roemer, *Political Competition* (Cambridge, MA: Harvard University Press, 2001).

# 12 Persons and situations

Shame and guilt, or contempt and anger, differ in that the first emotion in each pair targets a person's *character* and the second some *action* by the person (Chapter 8). Similarly, pridefulness rests on the belief that one is a superior person, and pride on the belief that one has performed some outstanding deed. But when we blame or praise an action, is it not because we believe it reflects the agent's character? To what other factor could it be ascribed?

### When folk psychology goes wrong

This book is mostly not about praise or blame, but about the *explanation* of behavior. In this context, the question is the power of character to explain action. People are often assumed to have personality traits (introvert, timid, etc.) as well as virtues (honesty, courage, etc.) or vices (the seven deadly sins, etc.). In folk psychology, these features are assumed to be stable over time and across situations. Proverbs in all languages testify to this assumption. "Who tells one lie will tell a hundred." "Who lies also steals." "Who steals an egg will steal an ox." "Who keeps faith in small matters, does so in large ones." "Who is caught red-handed once will always be distrusted."[1] If folk psychology is right, predicting and explaining behavior should be easy. A single action will reveal the underlying trait or disposition and allow us to predict behavior on an indefinite number of other occasions when the disposition could manifest itself. The procedure is not tautological, as it would be if we took cheating on an exam as evidence of dishonesty and then used the trait of dishonesty to explain the cheating. Instead, it amounts to using cheating on an exam as evidence for a trait (dishonesty) that will also cause the person to be unfaithful to a spouse. If one accepts the more extreme folk theory that all virtues go together, the cheating might also be used to predict cowardice in battle or excessive drinking.

[1] Presumably as a parody of such assertions, Thomas de Quincy wrote that if "once a man indulges himself in murder, very soon he comes to think little of robbing, and from robbing he comes next to drinking and Sabbath-breaking, and from that to incivility and procrastination."

People often make strong inferences from the austere private conduct of others. The British politician George Lansbury had a favorable impression of Hitler, based on the fact that "he has no love of show or pomp, is a total abstainer, non-smoker, vegetarian, and lives in the country rather than in a town. He is a bachelor, and likes children and old people." A member of the French Academy reportedly voted for de Gaulle because of the dignity of his private life, with the tacit premise that someone who would betray his wife is also likely to betray his country. In Vietnam, Communist leaders were able to win "the minds and the hearts of the population" because of their incorruptible personal style, in stark contrast to the less self-denying organizers from other political groups. Among the Mafiosi, having affairs is thought to be a sign of a disorderly and weak character.

Judging from the forensic speeches they left us, the ancient Greeks were strong believers in the unity and cross-situational consistency of character. A typical defense against an accusation was not "He did not do it," but "Given his excellent character as shown by his behavior on other occasions, he could not have done it." As one historian writes, "Witnesses are known to lie; they are not impartial observers, but in Euripides' words, competitors, who testify in support of the party that calls them. On the other hand, a man's deeds and associations are (at least in principle) known to the community in which he lives and are thought to reveal his true character." In discussing fourteen (!) arguments that were propagated at the time for the complicity of Queen Mary of Scots in the assassination of Lord Darnley, Hume cites one counterargument and its rebuttal: "That the only circumstance, which opposed all these presumptions or rather proofs, was, the benignity and goodness of her preceding behaviour, which seemed to remove her from all suspicions of such atrocious inhumanity; but that the characters of men were extremely variable, and persons, guilty of the worst actions, were not always naturally of the worst and most criminal dispositions." Although Hume does not say so, I conjecture that he endorsed the rebuttal.

To some extent, folk psychology is self-fulfilling. If people *believe* that others will predict their behavior in a situation of type A on the basis of their behavior in a situation of type B, they will act in situation B with situation A in mind. If the belief in a link between private and public morality is widespread (and known to be so), it creates an incentive for politicians to behave honestly in private life, assuming that any misbehavior would be made known to the electorate. Or suppose that it is widely believed that people have the same rate of time discounting in all situations. If they care too little about their future to take care of their bodies, they are also (according to folk psychology) likely to break a promise to realize a large short-term gain. Hence to be able to make credible promises about mutually profitable long-term cooperation, one should also cultivate a slim and healthy appearance.

To a larger extent, however, *folk psychology is demonstrably false*.[2] If one can eliminate the effects of folk psychology itself, so that there is no incentive to live up to expectations of cross-situational consistency, little consistency is found. Parents who only observe the behavior of their children at home are routinely surprised to learn that they are much more well behaved at school or when visiting the homes of classmates. Moreover, interventions to improve behavior in the family do not lead to improved adjustment at school, compared to that of control groups that received no intervention.[3] In laboratory experiments, most people (about two-thirds of the subjects) can be induced to behave heartlessly, to the point of imposing (what they believe to be) severe electrical shocks (about 450 volts) on a confederate of the experimenter. Yet there is no reason to believe that their behavior is due to an underlying trait of sadism, cruelty, or indifference to the suffering of others; in fact, many of the subjects who behaved in this way were upset and torn by what they were doing. Children are much more willing to wait for a larger delayed reward when both that reward and a smaller one that could be obtained immediately are hidden from sight. Any academic will know other academics who are conscientious in their research, but less so in their teaching or in their administrative tasks. Being talkative at lunch turns out to be poorly correlated with talkativeness on other occasions. A person may procrastinate in cleaning up the house but never on the job.[4]

In an essay "On the inconstancy of our actions," Montaigne contrasts the behavior of the younger Cato with that of ordinary humans such as he: "Strike one of [Cato's] keys and you have struck them all; there is in him a harmony of sounds in perfect concord such as no one can deny. In our cases on the contrary everyone of our actions requires to be judged on its own: the surest way in my opinion would be to refer each of them to its immediate circumstances, without looking farther and without drawing any firm inference from it." As he also notes, "if [a man] cannot bear slander but is resolute in poverty; if he cannot

[2] Not only folk psychology: economists who argue that agents signal through their behavior whether they are "good types" or "bad types" also overestimate consistency.

[3] There is a twist to these findings. The children of parents who complied with the advice of the interventionists did better at school than the children of non-compliant parents, a fact cited by some in support of the spillover from home to school. Yet the finding may be due simply to the fact that compliance is inheritable. Parents who conscientiously follow the instructions of an interventionist are more likely to have children who conscientiously follow the instructions of a teacher.

[4] In a letter to the ethicist Randy Cohen (*New York Times Magazine*, January 15, 2006) an academic asked whether the fact that an untenured colleague claimed discounts at the Faculty Club to which he was not entitled warranted a vote against him for tenure "because of his dishonesty and its potential extension to his research." Cohen said no, on the grounds that "people who behave badly in some situations often behave well in others."

bear a barber-surgeon's lancet but is unyielding against the swords of his adversaries, then it is not the man who deserves praise but the deed."[5]

According to some scholars, *being overweight* signals bad self-control (or a high rate of time discounting) and allows others to predict that the person will also be unable to keep his promises or engage successfully in long-term ventures. Counterexamples come easily to mind. Louis XVIII was grossly overweight. On one occasion, he consumed 180 oysters at one sitting. Yet he also persevered unflinchingly in his efforts to regain the throne of France during twenty-five years of exile. More recently, Governor Christie of New Jersey

> addressed the fact that there has been discussion over his weight in recent weeks, saying the jokes did not bother him. "I'm not particularly self-conscious about this," he said, adding: "It's not a news flash to me that I'm overweight." While Christie said he found many of the jokes about his weight funny, Americans should "look down upon" the "people who pretend to be serious commentators" who suggested he couldn't be president because of his weight. "The people who wrote [that] are ignorant people," said Christie. "To say that because you are overweight you are therefore undisciplined – I don't think undisciplined people get to achieve great positions in our society," he said (*Washington Post*, October 5, 2011).

Let me give some examples from art and artists. Proust wrote that "one might have thought" that the young men in *Le temps retrouvé* who were paid for inflicting pain on the customers of Jupien's brothel must be "fundamentally bad, but not only were they wonderful soldiers during the war, true 'heroes,' they had just as often been kind and generous in civil life." Commenting on the apparent contradiction between Swann's "exquisite dissimulation of an invitation to Buckingham Palace" and his boast that the wife of a lower functionary had visited Mme Swann, he wrote that

> the main reason was (and this is one that holds for all of humanity) that even our virtues are not extraneous, free-floating things which are always at our disposal; in fact they come to be so closely linked in our minds with the occasions for acting on which we feel they should be deployed that, if we are required to engage in some different activity, it can take us by surprise, so that we never even think that it too might entail the use of those very virtues.

The jazz musician Charlie Parker was characterized by a doctor who knew him as "a man living from moment to moment. A man living for the pleasure principle, music, food, sex, drugs, kicks, his personality [*sic*] arrested at an infantile level." Another great jazz musician, Django Reinhardt, had an even

[5] Yet he also said, "no man who has been a real fool once will ever be really wise again." Is it true? In Western societies some former Stalinists and Maoists have joined the community of reason, but others have not. Some went from one kind of foolishness to another.

more extreme present-oriented attitude in his daily life, never saving any of his substantial earnings, but spending them on whims or on expensive cars, which he quickly proceeded to crash. In many ways he was the incarnation of the stereotype of "the Gypsy." Yet you do not become a musician of the caliber of Parker and Reinhardt if you live in the moment *in all respects*. Proficiency takes years of utter dedication and concentration. In Reinhardt's case, this was dramatically brought out when he damaged his left hand severely in a fire and retrained himself so that he could achieve more with two fingers than anyone else with four. If these two musicians had been impulsive and carefree across the board – if their "personality" had been consistently "infantile" – they could never have become such consummate artists.

After 1945, the Norwegian novelist Knut Hamsun, who had collaborated with the Nazis during the war, underwent psychiatric observation to determine whether he was mentally capable of being tried (he was eighty-six years old at the time). When the psychiatric professor asked him to describe his "main character traits," he replied as follows:

> The so-called naturalistic period – Zola and his time – wrote about persons with main character traits. They had no use for nuanced psychology. People had one dominant capacity that governed their actions. Dostoyevsky and others taught us all something different about people. From the very beginning I do not think there is a single person in any of my writings with this dominant and unitary capacity. They are all without so-called character – they are divided and fragmented, not good not bad, but both. Nuanced and changing in their mind and in their actions. This is no doubt how I am myself. It is very possible that I am aggressive, and that I have a little of the other traits the professor suggested – vulnerable, suspicious, selfish, generous, jealous, righteous, logical, sensitive, a cold nature. All these would be human traits, but I cannot give any of them the preponderance in myself.

In Chapter 16, I pursue the question of character or "lack of character" in works of fiction. Here I shall only note that Hamsun does not refer to the possibility that he might be, for instance, consistently generous in one type of situation and consistently selfish in another. I now turn to this issue.

## The power of the situation

"Being impulsive with money" and "being dedicated to one's music," "being talkative at lunch," or "being conscientious in one's research" are also, of course, character traits. They are, however, situation-specific or *local traits* rather than global personality features that manifest themselves across the board in all situations. Contrary to folk psychology, systematic studies find very low levels of cross-situational consistency for character traits. Although correlations exist, they are typically so low that they cannot be detected "by the naked eye." Psychopaths may exhibit uncaring behavior across the

board,[6] and the younger Cato may have been consistently heroic, but for the great majority of individuals who fall between these extremes such consistency is not to be expected. The more extreme idea of folk psychology, according to which all virtues go together, has not been as thoroughly tested, perhaps because it seems so obviously implausible. Yet it may still have a grip on the mind, as shown by our confidence in the medical skills of doctors who have good "bedside manners." In classical antiquity, the idea that excellence in one arena was an infallible predictor or "index" of excellence in others was common. Psychologists refer to this as a "halo effect."

Often, therefore, the explanation of behavior is found in the *situation* rather than in the *person*. Consider for instance the fact that some Germans acted to rescue Jews from the Nazi regime. On a "characterological" theory, one would assume that the rescuers had an altruistic personality type that non-rescuers lacked. It turns out, however, that the factor with the strongest explanatory power was the "situational" fact of *being asked* to rescue someone. The causal link could arise in two ways. On the one hand, it is only by being asked that one can obtain the *information* that is needed to act as a rescuer. On the other hand, the face-to-face situation of being asked might trigger acceptance because of the *shame* one would feel if one refused.[7] The first explanation assumes altruism but denies that it is sufficient to explain the behavior. The second denies altruism and substitutes social norms for moral norms. On either account, what differentiates rescuers from non-rescuers is the situation in which they find themselves rather than their personality.

The "Kitty Genovese" case is another example of the power of the situation. It is implausible to stipulate, on the basis of their inaction, that all the witnesses to her murder were callous and indifferent to human suffering. Rather, many of them may have thought that someone else was going to call the police, or that since nobody was doing anything about it the situation was not as serious as it might seem ("probably just a domestic dispute"), or that the inaction of the others suggested that direct intervention might be risky.[8] These lines of reasoning become more plausible the greater the number of passive

[6] Since an intelligent egoist who cared about the future would often have an interest in *mimicking* concern for others (Chapter 5), the ultimate explanation of psychopathic behavior might be excessive discounting of the future.

[7] Similarly, the success of telethons in making people give money does not rest on their appeal to altruistic motives but to the fact that they are accompanied by a knock on the door by someone making a face-to-face request. In this case, the information-based explanation is clearly inadequate.

[8] People who were afraid of intervening physically to protect the victim from her assailant might still have called the police. At the time, however, the police did not accept anonymous calls, so that bystanders might have been afraid of getting into trouble. In other situations of this kind the option of calling the police may be unavailable.

bystanders. Thus in one experiment, subjects heard a confederate of the experimenter feigning an epileptic seizure over the intercom system. When subjects believed they were the only listener, 85 percent intervened to help; when they believed there was one other listener, 62 percent intervened; when they believed there were four others, 31 percent intervened. In another experiment, 70 percent of lone bystanders intervened but only 7 percent did so when sitting next to an impassive confederate.[9] With two naive subjects, the victim received help in 40 percent of the cases. Thus not only does the chance that any *given* bystander will intervene go down when there are more of them, but the chance that *some* bystander will intervene also falls with the number of bystanders. In other words, the dilution of the responsibility to intervene caused by the presence of others occurs so fast that it cannot be offset by the greater number of potential interveners.

In another experiment, theology students were told to prepare themselves to give a brief talk in a nearby building. One-half were told to build the talk around the Good Samaritan parable (!), whereas the others were given a more neutral topic. One group was told to hurry since the people in the other building were waiting for them, whereas another was told that they had plenty of time. On their way to the other building, subjects came upon a man slumping in a doorway, apparently in distress. Among the students who were told they were late, only 10 percent offered assistance; in the other group, 63 percent did so. The group that had been told to prepare a talk on the Good Samaritan was not more likely to behave as one. Nor was the behavior of the students correlated with answers to a questionnaire intended to measure whether their interest in religion was due to the desire for personal salvation or to a desire to help others.[10] The situational factor – being hurried or not – had much greater explanatory power than any dispositional factor.

It would not be accurate to subsume this analysis under that of the previous chapter, by saying that the students in the "hurry" category behaved the way they did because of a *time constraint*. Their constraint was not an objective or "hard" one, and in fact 10 percent of the students in this group did offer assistance. Rather, the situation shaped behavior by affecting the salience of competing *desires*. The face-to-face request enhances the strength of

[9] This is at least the general tendency in the numerous experiments of this kind that have been carried out. In the one just cited (the epileptic seizure heard over the intercom) it turns out that if we assume that the other listeners were real, naive subjects who received the same information (and not simply confederates or fictions created by the experimenter), the chance that at least one of them would intervene is roughly constant, that is, around 85 percent. In the case of five subjects (the main subject and the four listeners), the chance that any one of them would abstain from intervening is 0.69. The chance that all of them would abstain is $(0.69)^5$, or 0.156, yielding a likelihood of 0.844 that at least one would intervene.

[10] Just like the subjects who were induced to inflict electrical shocks, many of those who hurried by the man in distress were themselves visibly distressed by the encounter.

other-regarding motives, whereas being told to hurry diminishes it. Being able to *see* the reward that is imminently available makes it more attractive compared to one that will only be available with a delay, just as the sight of a beggar in the street can trigger generosity that the abstract knowledge of poverty would not. "Kitty Genovese" situations change both the perceived costs and perceived benefits of helping. The desire to comply with instructions by an impassive experimenter that "you must continue" to administer apparently painful and possibly fatal electrical shocks overrules the desire not to inflict pain needlessly.

There is no general or common mechanism by which a situation can affect behavior. Situations range from face-to-face demands to rescue Jews to the most trivial events, such as when finding a quarter in the coin return slot of a pay phone lifts one's mood and makes one help a stranger (in reality a confederate of the experimenter) retrieve a bunch of papers dropped on the sidewalk. The important lesson from these observations, in real life and in the laboratory, is merely that *behavior is often no more stable than the situations that shape it*. A person may be talkative at lunch when he can relax with long-standing colleagues and be tongue-tied with strangers. A person may consistently give to beggars but otherwise not give a thought to the poor. A person may invariably be helpful in situations in which nobody else can help, and invariably passive in the presence of other potential helpers. A man may be consistently aggressive and make biting remarks to his wife, yet be calm and generous to other people. His wife, too, may display the same dual behavior. His aggression triggers hers, and vice versa.[11] If they rarely see the spouse interact with other adults, for example, at the workplace, they may believe that he or she is intrinsically aggressive rather than merely aggressive in the situation defined by their presence.

### The spontaneous appeal to dispositions

To pursue the last example, marital therapists often try to make the spouses who seek their help switch from character language to action language (see Chapter 8). Rather than saying, "You are a bad person," thereby leaving little room for hope or change, they should make an effort to say, "You did a bad thing." The latter phrasing leaves open the possibility that the action in question might have been triggered by specific situational factors, such as a provocative remark by the other spouse. One reason (among many) why therapists often have little success in reframing conflicts in this way is that

[11] Metaphorically speaking, they are in a "bad psychological equilibrium." Yet aggression need not be a "best response" to aggression (as required by the game-theoretic notion of equilibrium), only a psychologically intelligible one.

people spontaneously privilege character-based explanations of behavior over situation-based ones. If we learn that somebody has contributed to a "gay rights" ad, we tend to assume that the person *is* gay or liberal rather than that he *was asked* in a way that made it hard to refuse. When interviewing a job candidate, we tend to explain what the person says or does in light of the dispositions we (overconfidently) impute to him or her, rather than to the special nature of the interview situation. Language itself reflects the dispositional bias. Adjectives that apply to actions ("hostile," "selfish," or "aggressive") can usually also be applied to the agent, whereas there are few characterizations of actions that also apply to situations ("difficult" is an exception).

Psychologists refer to the inappropriate use of dispositional explanation as the *fundamental attribution error*, that is, explaining situation-induced behavior as caused by enduring character traits of the agent. When subjects were asked to predict the behavior of the theology students who encountered a distressed individual, they (wrongly) thought that people whose religion was based on a desire to help others would be more likely to act as a Good Samaritan and (again wrongly) that being in a greater or lesser hurry would make no difference at all. Other subjects overpredicted infliction of electrical shocks in the absence of the specific situational factors in the original experiment, thus revealing their belief in a dispositional explanation. When an instructor assigns a student the task of writing a pro-Castro essay, other students, knowing how it was assigned, still interpret the essay as manifesting a pro-Castro attitude. When students were asked to volunteer for tasks with either low or high remuneration, and low or high numbers of volunteers resulted, observers, knowing the pay differential, nevertheless predicted that *all* volunteers were more likely than non-volunteers to volunteer for a non-paying cause. The observers, in other words, attributed the action of volunteering to a disposition to volunteer rather than to the reward structure of the situation.

People in some societies seem less prone than those in others to the fundamental attribution error. Experiments indicate that compared to Americans, Asians ascribe more importance to the situation and less to personal dispositions in explaining behavior. Real-life situations, too, display this difference. Thus in 1991, an unsuccessful Chinese physics student shot his adviser, several fellow students, and then himself. In the same year, an American postal employee who had lost his job shot his supervisor, several fellow workers and bystanders, and then himself. Both events were widely reported in English and in Chinese newspapers, the former consistently explaining them in dispositional terms ("disturbed," "bad temper," "mentally unstable") and the latter in situational terms ("easy access to guns," "had just lost his job," "victim of pressure to succeed"). Other findings confirm this

difference. It might be due, however, to the fact that situational factors actually play a greater role in generating the behavior of Asians. Rather than being better at overcoming the dispositional bias, they may have less of a bias to overcome, or both of these factors might operate.

Overcoming the fundamental attribution error can be liberating. First-year college students who are told that most freshmen do poorly but that their grades subsequently improve, in fact do somewhat better in later years than those who are not given this information. The latter are more likely to impute their poor performance to their low ability than to the unfamiliar and distracting college environment. Not believing they can do better, they are less motivated to try. When oppressed groups shed the essentialism of their oppressors – the idea that women, blacks, or Jews are intrinsically inferior – they can more easily shed their shackles.

Is the fundamental attribution error hot or cold – a motivated mistake or more along the lines of an optical illusion? To the extent that motivation enters into the process of attribution, there is no reason why it should consistently lead us to overemphasize dispositions. On self-serving grounds, we should attribute our success to our enduring character traits, and our failures to unfortunate circumstances.[12] If the French moralists are to be believed, we should attribute the successes of others to their good luck and their failures to their dispositions.[13] On cognitive grounds, the tendency to favor the person over the situation may be an instance of a more general tendency to pay more attention to the moving foreground than to the static background. It follows that the error should be less common in cultures in which more evenhanded attention is paid to foreground and background, as seems to be the case in Asian cultures.

### The rehabilitation of the person

The findings I have described undermine what one might call "crude essentialism" in the study of personality. It is simply not true that people *are* aggressive, impatient, extroverted, or talkative across the board. At the same time, the findings do not imply that the situation is all-powerful in explaining behavior. Rather, we have to decompose "the" character into a set of *contingent* response tendencies. Instead of characterizing a person as altruistic, we might describe

[12] Sometimes, though, we may be motivated to impute our failures to our character. A gambler or an alcoholic may be happy to tell himself that he "just cannot help it" in order to have an excuse for persisting. As I noted in Chapter 9, students may claim that they are unable to do their homework when in reality they are just unwilling.

[13] Sometimes, though, we may be motivated to impute the successes of others to their character *flaws*. Anti-Semitism relies on the myth that Jews succeed because their immoral character makes them willing to adopt any means to get ahead.

him or her by the phrase "helps when asked, but does not volunteer to help," or by the phrase "helps when unstressed, but is neglectful when stressed." Each of these phrases might characterize one aspect of a person and thus underwrite a more subtle form of essentialism. A person might scold a spouse for never cleaning up around the house ("you're lazy") or for never cleaning up unless asked to ("you're thoughtless"). In the latter case, the spouse might be proactive rather than reactive in other matters, such as monitoring the health of the children in the family. There would be no across-the-board trait of reactiveness.

In this perspective, explanation of behavior rests on the particular situation plus the person-specific relation between situations and behavioral propensity. One person might be highly aggressive with individuals over whom he has power, but exceptionally friendly with those who have power over him, whereas another person might show the opposite pattern. If we observe both of them behaving in a friendly manner, we might be tempted to conclude that they both *are* of a friendly disposition. As should be clear by now, however, the similarity of behavior might be due to differences in situation and in response contingencies that exactly cancel each other.

### Bibliographical note

The "tendency to overestimate the unity of personality" was clearly stated by G. Ichheiser, "Misunderstandings in human relations: a study in false social perception," *American Journal of Sociology* 55 (1949), Supplement. Recent work deemphasizing "character" derives from W. Mischel, *Personality and Assessment* (New York: Wiley, 1968). The present exposition relies heavily on L. Ross and R. Nisbett, *The Person and the Situation* (Philadelphia: Temple University Press, 1991), and on J. Doris, *Lack of Character* (Cambridge University Press, 2002). George Lansbury's comment on Hitler is in his book *The Quest for Peace* (London: Michael Joseph, 1938), p. 141. The references to Communist organizers in Vietnam and to Mafiosi are from, respectively, S. Popkin, *The Rational Peasant* (Berkeley: University of California Press, 1979), and D. Gambetta, "Trust's odd ways," in J. Elster *et al.* (eds.), *Understanding Choice, Explaining Behavior: Essays in Honour of Ole-Jørgen Skog* (Oslo: Academic Press, 2006). The reference to the effects of intervention is from J. R. Harris, *No Two Alike* (New York: Norton, 2006). The willingness to inflict electrical shocks is described in a classic study by S. Milgram, *Obedience to Authority* (New York: Harper, 1983). A useful perspective on his account is G. Perry, *Behind the Shock Machine* (New York: The New Press, 2012). The information about the two musicians is found in R. Russell, *Bird Lives: The High Life and Hard Times of Charlie (Yardbird) Parker* (New York: Charterhouse, 1973), and M. Dregni, *Django: The Life and Music of a*

*Gypsy Legend* (Oxford University Press, 2004). The statement by Hamsun is translated from G. Langfeldt and Ø. Ødegård, *Den rettspsykiatriske erklæringen om Knut Hamsun* (Oslo: Gyldendal, 1978), p. 82. The use of behavior as an "index" in antiquity is discussed by P. Veyne, *Le pain et le cirque* (Paris: Seuil, 1976), pp. 114, 773; see also P. Veyne, "Pourquoi veut-on qu'un prince ait des vertus privées?" *Social Science Information* 37 (1998), 407–15. The "characterological" explanation of the willingness to rescue Jews is argued by K. Monroe, M. C. Barton, and U. Klingemann, "Altruism and the theory of rational action: rescuers of Jews in Nazi Europe," *Ethics* 101 (1990), 103–22. The "situationist" explanation is argued by F. Varese and M. Yaish, "The importance of being asked: the rescue of Jews in Nazi Europe," *Rationality and Society* 12 (2000), 307–24. A skeptical note about the tendency to infer dispositions from behavior is sounded in J. L. Hilton, S. Fein, and D. Miller, "Suspicion and dispositional inference," *Personality and Social Psychology Bulletin* 19 (1993), 501–12. The contrast between Americans and Asians is summarized in R. Nisbett, *The Geography of Thought* (New York: Free Press, 2004). What I call the "rehabilitation of the person" is argued in W. Mischel, "Towards an integrative science of the person," *Annual Review of Psychology* 55 (2004), 1–22.

# 13 Rational choice

## The structure of rational action

In this chapter, and in Chapter 18, I state some normative principles about how people should behave to realize their aims, whatever they might be, as well as possible. These principles can also have explanatory power, if we assume that social agents abide by them. In Chapter 14 and Chapter 19 I confront this assumption with empirical observations, and find it to be partly unfounded.

Rational-choice theorists want to explain behavior on the bare assumption that agents are rational. This assumption includes the hypothesis that agents form rational beliefs, including beliefs about the options available to them. There is no need, therefore, to classify the determinants of behavior as either subjective (desires) or objective (opportunities). Rational-choice theory is subjective through and through.

The structure of rational-choice explanation is laid out in Figure 13.1. An action is rational, in this scheme, if it meets *three optimality requirements*: the action must be optimal, given the beliefs; the beliefs must be as well supported as possible, given the evidence; and the evidence must result from an optimal investment in information gathering. In Figure 13.1 the arrows have a double interpretation, in terms of causality as well as of optimality. The action, for instance, should be caused by the desires and beliefs that make it a rational one; it is not enough to do the right thing by fluke. Similarly, a belief is not rational if it is the outcome of two oppositely biased processes that exactly cancel each other. To take an example, smokers as well as non-smokers process information about the dangers of smoking in ways that make them believe these are greater than in fact they are. At the same time, smokers are subject to a self-serving bias that makes them discount the risks. If as a result they form the same belief as an unbiased observer would hold,[1] that does not prove they are rational. In one of the most influential discussions of rationality in the social sciences, Max Weber made the mistake of inferring "process rationality" from "outcome optimality" when he wrote that:

[1] As a matter of fact, the second bias of the smokers does not fully compensate for the first.

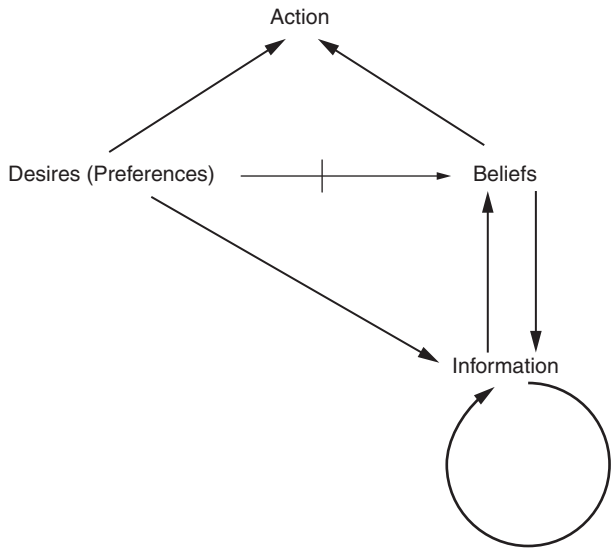



Figure 13.1

for the purposes of a typological scientific analysis it is convenient to treat all irrational, affectually determined elements of behavior as factors of deviation from a conceptually pure type of rational action. For example a panic on the stock exchange can be most conveniently analyzed by attempting to determine first what the course of action would have been had it not been influenced by irrational affects; it is then possible to introduce the irrational components as accounting for the observed deviations from this hypothetical course. Similarly, in analyzing a political or military campaign it is convenient to determine in the first place what would have been a rational course, given the ends of the participants and adequate knowledge of all the circumstances. *Only in this* way is it possible to assess the causal significance of irrational factors as accounting for the deviation from this type.

Although Weber was right in thinking that deviation from the rational course of action is a sufficient condition for irrationality to be at work, he erred in asserting (in the phrase I have italicized) that it was a necessary one. A similar mistake is involved in asserting that instinctive fear reactions are rational, when all that can be said is that they are *adaptive*. When I see a shape on the path that may be either a stick or a snake, it makes sense to run away immediately rather than gathering more information. It seems that human beings are in fact hardwired to do so. This flight behavior is not rational in the strict sense, since it is not produced by the machinery of rational decision making, yet it mimics rationality in the sense of being the very same behavior that that machinery would have produced had it been brought to bear on the

situation. When the opportunity costs of gathering information (Chapter 14) are high, a rational agent will not collect much of it. Yet often the flight tendency is not caused by such calculations, but preempts them.[2]

### Preferences and ordinal utility

Spelled out more fully, the first optimality requirement is that the action must be the best means of satisfying the agent's desires, given his beliefs about the available options and their consequences. What is "best" is defined in terms of "betterness" or preference: the best is that than which none is better, as judged by the agent. There is no implication that the desires be *selfish*. The confusion of rationality and egoism is a crude error, although one that is facilitated by the practice of some rational-choice theorists. Nor do we need to require that desires be *stable*, not even in the minimal sense of excluding temporary preference changes. An agent who under the influence of emotion or drugs prefers A to B acts rationally in choosing A, even if she under other circumstances prefers B to A. A case in point (see Chapter 6) occurs when the weight the agent assigns to future consequences of present choice is diminished as the result of such influences.

For the analysis to get off the ground, the notion of "best" has to be well defined. Barring technicalities, two conditions ensure that this will be the case. First, preferences have to be *transitive*. Suppose there are three options, A, B, and C. If a person thinks A is at least as good as B and B at least as good as C, he should also think A at least as good as C. If transitivity fails, for instance if the person strictly prefers A to B, B to C, and C to A, he may not have a "best" option. Moreover another person can exploit this fact, by offering the agent a move from a less preferred to a more preferred option in return for a sum of money. Since preferences cycle, this operation can be repeated indefinitely, bringing about the person's ruin by a series of stepwise improvements.[3]

This situation can arise if the agent ranks the options by "counting aspects." Suppose I choose one apple over another if it is better in at least two out of three aspects, such as price, taste, and perishability. If apple A beats apple B in price and taste, apple B beats apple C in price and perishability, and apple C beats apple A in taste and perishability, transitivity is violated. Although this possibility is relatively unimportant in individual choice, in which it merely reflects the failure of a rule of thumb, we shall see (Chapter 24) that it is more significant in collective choice.

[2] In rats, the delay between the unthinking response and the reflective one is about 10 milliseconds.

[3] An agent who discounts the future hyperbolically may also be trapped in this way. See Chapter 9 for another way of "improving oneself to death."

A different problem arises when indifference fails to be transitive. I may be indifferent between A and B and between B and C, because the differences within each pair are too small to be noticeable, but prefer C to A because there is a detectable difference between them. There is an option that is "best," namely, C, but it is still possible to make the agent worse off by making her a series of offers – exchanging C for B and B for A – that she has no reason to refuse and hence might well accept. What justifies calling an agent with intransitive preferences irrational is not so much the lack of a "best" option, but the fact that she may accept offers that make her worse off.

To ensure that the idea of "the best" is always a meaningful one we must also require that preferences be *complete*: for any two outcomes the agent should be able say whether he prefers the first to the second, prefers the second to the first, or is indifferent between them. If he is unable to make any of these three responses, he may not be able to determine which option is the best. I say more about incompleteness toward the end of the chapter. Here, I only want to note that unlike lack of transitivity, a lack of completeness is not any kind of failure. Suppose I want to give an ice cream to the one of two children who will enjoy it most. For *me* to have a preference over the two options, I would have to be able to compare *their* levels of preference satisfaction were they given the ice cream. Often, however, this is an impossible task. The failure to carry it out is not a failure, in the sense that I could have done better, but reflects simply a fact of life.

For many purposes, transitivity and completeness of preferences are all we need to identify the rational action. It is often convenient, however, to represent preferences by numbers, often called *utility values*, that are assigned to the options. To ensure this possibility we impose a further condition on preferences: *continuity*. If each option in a sequence A1, A2, A3, . . ., is preferred to B and the sequence converges to A, then A should be preferred to B; if B is preferred to each option in the sequence, B should be preferred to A. A counterexample is provided by "lexicographic preferences": a bundle of two goods A and B in quantities (A1, B1) is preferred to another bundle (A2, B2) if and only if either A1 > A2 or (A1 = A2 and B1 > B2). In this preference ranking, the bundles (1.1, 1), (1.01, 1), (1.001, 1), . . ., are all preferred to (1, 2), which is preferred to (1, 1). Loosely speaking, we may say that the first component of the bundle is incomparably more important than the second, since no extra amount of good B can offset even the smallest loss of good A.[4] Or, more simply, no trade-off is possible. Hence these preferences cannot be represented by indifference curves. Whereas lexicographic preferences rarely if ever apply to ordinary consumption goods, they can matter for

[4] The intuitive notion of incomparability may, therefore, be spelled out in two distinct ways: as incomplete preferences or as discontinuous preferences.

political choices. A voter may prefer candidate A to candidate B if and only if A has a stronger pro-life attitude on abortion *or* if they have the same attitude on that issue and A proposes lower taxes than does B. For such voters, the "sacred value" of life may not be traded off against the secular value of money.

If the agent's preferences are complete, transitive, and continuous, we can represent them by a continuous utility function u that assigns a number u(A) to each option (A). Instead of saying that a rational agent chooses the best feasible option, we may then say that the agent *maximizes utility*. In this phrase, "utility" is a mere shorthand for preferences with certain properties. To see this, we may note that the only requirement for a function u to represent a preference order is that A is preferred to B if and only if $u(A) > u(B)$. If u is always positive, $v = u^2$ can also represent the same preference order, although v assigns larger or (for $u < 1$) smaller numbers than u. The absolute numbers have no significance; only their relative or *ordinal* magnitude has. Hence the idea of "utility-maximization" does not imply that the agent is engaged in getting as much as possible of some psychic "stuff." It does, however, exclude the kind of value hierarchy embodied in lexicographic preferences. These cannot, in fact, be represented by a utility function.

### Cardinal utility and risk attitudes

Often, agents face *risky* options, that is, choices that may, with known probabilities, have more than one possible outcome. Intuitively, it would seem that a rational agent would choose the option with the greatest *expected utility*, an idea that incorporates the utility of each outcome as well as its probability of occurrence. She would first, for each option, weigh the utility of each consequence by its probability and add up all the weighted utilities, and then choose the option with the greatest sum.

Ordinal utility does *not* allow us, however, to spell out this idea. Suppose there are two options, A and B. A can produce outcome O1 or O2 with probabilities 1/2 and 1/2, whereas B can produce outcome O3 or O4 with probabilities 1/2 and 1/2. Assume now a utility function u that assigns values 3, 4, 1, and 5 to O1, O2, O3, O4, respectively. The "expected ordinal utility" of A is 3.5 and that of B is 3. If instead we use the function $v = u^2$, the numbers are 12.5 and 13. Each function represents preferences as well as the other, and yet they single out different options as "the best." Clearly, this approach is useless.

It is possible to do better, but at some conceptual costs. The approach associated with John von Neumann and Oskar Morgenstern shows that one can assign the options utility values that have a *cardinal* and not merely ordinal significance. An instance of a cardinal value assignment is temperature. Whether we measure temperature in Celsius or Fahrenheit does not affect the truth value of the statement "the average temperature in Paris is higher than the average

temperature in New York." (If temperatures were measured ordinally, this statement would not make sense.) By contrast, the truth value of the statement "It is twice as hot in Paris as in New York" *does* depend on the choice of scale. Yet although the truth value of this particular statement about intensities is scale sensitive, others are not. The truth value of the statement "The temperature difference between New York and Paris is greater than that between Paris and Oslo," for instance, does not depend on the choice of scale. Similarly, we can construct cardinal measures of utility that reflect – among other things, as we shall see – the intensity of preferences and not merely the ordinal ranking of options. These enable us to compare the utility gain (or loss) of going from x to (x + 1) to that of going from (x + 1) to (x + 2), that is, to talk about increasing or decreasing marginal utility – concepts that are meaningless for ordinal utility measures.

The technical details of the construction need not concern us, as the basic idea is simple and sufficient for present purposes. We begin by assuming that agents have preferences not simply over options, but over *lotteries* of options (including the "degenerate lotteries" that consist of getting a basic option for sure). For any given set of basic options or "prizes," a lottery specifies, for each prize, the probability of obtaining it, the probabilities adding up to 1. Agents are assumed to have complete and transitive preferences over such lotteries. Preferences are also assumed to obey an "independence axiom": the preference between two lotteries p and q is unaffected if they are both combined in the same way with a third lottery r. The "certainty effect" cited in Chapter 7 and further discussed in Chapter 14 violates this axiom.

Finally, preferences are assumed to exhibit a form of continuity, defined as follows. Suppose the basic options include a best element A and a worst element B. We assign them, arbitrarily, utility numbers 1 and 0. Continuity means that for any intermediate option C there is a probability p(C) that would make the agent indifferent between getting C for certain and engaging in a lottery that would give him A with probability p(C) and B with probability 1 – p(C).[5] We then define the *cardinal utility* u(C) as equal to p(C). This number, to be sure, is arbitrary because the end-point utilities are. Suppose we assign utility numbers M and N to A and B, respectively (M > N). We then define the utility of C as the expected utility of the lottery:

$$pM + (1 - p)N = Mp + N - Np = (M - N)p + N.$$

The class of utility functions that arise in this way is much smaller than the class of ordinal utility functions.[6] It is easy to see that if option X has greater

[5] *Identifying* this probability raises the problems of anchoring cited in the introduction to Part II.

[6] Any two such functions are in fact related to each other as are the Celsius and Fahrenheit temperature scales, which assign different values (corresponding to M and N in the text) to the temperatures at which water boils and freezes.

expected utility than Y according to one function, it will also have greater expected utility according to any other. Thus we can assert, without ambiguity, that a rational agent maximizes expected utility.

Cardinal utility functions have the important property of being *linear in probability*. Let us introduce the notation XpY, meaning a lottery that offers probability p of getting X and 1 – p of getting Y. Using the 1 – 0 end-point scale, the utility u(X) equals the probability q at which the agent is indifferent between X and the lottery AqB. Similarly, the utility u(Y) equals the probability r at which he is indifferent between Y and the lottery ArB. XpY, therefore, offers the utility equivalent of a chance p of getting A with probability q and a chance 1 – p of getting A with probability r. The utility of XpY, therefore, is pq + r(1 – p), which is p times the utility of X plus (1 – p) times the utility of Y. For instance, the utility of the probabilistic combination of a 3/5 chance of getting X and a 2/5 chance of getting Y is 3/5q + 2/5r.

Somebody could make the following objection. Suppose a farmer has the choice between two crops: the traditional variety that is equally likely to produce a good or a mediocre harvest, depending on the weather, and a modern variety that is equally likely to produce an excellent crop or a poor one. Suppose the cardinal utilities are 3 and 2 for the old crop, 5 or 1 for the new one. Since the expected utility of the new crop is larger, that is what the farmer ought to choose. But – the objection might go – does this not disregard the fact that the farmer might be risk-averse and unwilling to accept any option that might lead to a utility level as low as 1? The objection involves double-counting, however, as risk aversion is *already* incorporated in the construction of the cardinal utilities. Assuming that A, B, and C take the values of 100, 0, and 60, u(C) might well be 0.75 for a risk-averse person, implying that she is indifferent between getting 60 for certain and a lottery that leaves her with a 25 percent chance of getting nothing and a 75 percent chance of getting 100. A similar argument applies to the assignment of cardinal utility values to physical amounts of the crop.

For another illustration, consider the allocation of child custody (see Figure 13.2). The horizontal axis can be understood in two ways, as involving either a physical division of custody (percentage of the time spent with the child) or a probabilistic division (the chance of being awarded full custody in a court of law). The cardinal utility of equal time sharing is AE, which is greater than the utility AC of a 50 percent chance of full custody. (Here we appeal to the fact that cardinal utility is linear in probability.) The reason is that most people in this situation display risk aversion. They are willing to accept joint custody because a 50 percent risk of not being able to see the child at all is intolerable. It is only if a parent believes that his or her chance of getting full custody is greater than q percent that litigation is preferable to joint custody. If there is a considerable amount of custody litigation it is not because parents are

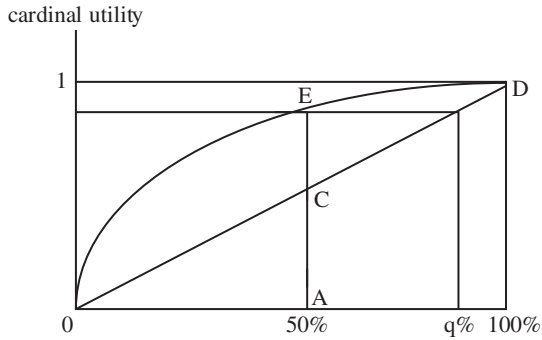



Figure 13.2

risk lovers, but because wishful thinking makes them exaggerate their chance of being awarded custody.

### Risk aversion and decreasing marginal utility

The preceding exposition, while accurate, could be misleading. There is a tendency in part of the literature to blur the distinction between risk aversion and decreasing marginal utility. To develop this point, I need to introduce a concept that is intuitively meaningful, although it has not (so far) lent itself to measurement. This is the idea of the *intrinsic utility* of a good, reflecting the intensity of preferences of the agent. Introspection tells us compellingly that some goods or experiences are immensely enjoyable, others merely satisfying, still others mildly annoying, and some downright dreadful. To represent the difference between them merely in terms of ordinal preferences – "I prefer heaven to hell, just as I prefer four apples to three" – is clearly to use a very impoverished notion of welfare or utility. The fact that there is no reliable way of assigning numbers to intrinsic levels of satisfaction or dissatisfaction does not prove that the idea is meaningless, any more than our inability to quantify and compare the levels of satisfaction of different individuals shows that the idea of interpersonal comparison of welfare is meaningless.

The idea that many goods have decreasing marginal utility may be understood in this perspective. For a poor person, the first dollars have great utility, but then each successive extra dollar becomes worth less in subjective terms. Every smoker knows that the first cigarette in the morning is the best one, and that you enjoy each cigarette more if you pace yourself and do not smoke too frequently. Smoking a cigarette, in fact, has two effects: producing enjoyment in the present and reducing the enjoyment of future cigarettes.

The second effect does not, however, have to be negative. Consider again the child custody case. For a parent, one afternoon with the child every other

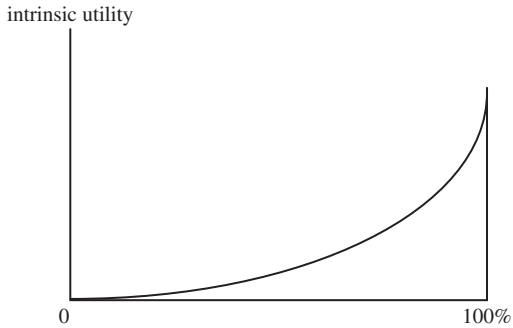



Figure 13.3

weekend may provide more frustration than satisfaction. An afternoon every weekend is more than twice as satisfying, because the stronger emotional bonds created by more frequent encounters make each of them more satisfying. At the other end of the time spectrum, the extra satisfaction of being with the child seven days a week rather than six exceeds the extra satisfaction of six days rather than five, because full custody provides the benefits of unconstrained planning. Being with the child, in fact, has increasing marginal (intrinsic) utility, as shown in Figure 13.3.

Here, the interpretation of the horizontal axis is the percentage of the time spent with the child. For the reasons just given, each extra hour is more valuable than the preceding one. *This statement is perfectly compatible with the analysis underlying* Figure 13.2. The marginal utility of time spent with the child may be decreasing if utility is understood as cardinal utility, but increasing if it is understood as intrinsic utility. The fact that only the first part of this statement has a measurable interpretation does not imply that the second is meaningless.

While cardinal utility functions are always generated by two underlying psychological factors, risk attitudes and intrinsic utility, these cannot be measured separately. We cannot tell in any rigorous way whether the curve OED in Figure 13.2 is derived from risk neutrality combined with decreasing marginal intrinsic utility of time spent with the child or from risk aversion combined with increasing marginal intrinsic utility of time spent with the child. In a given case, intuition may tell us that the one or the other interpretation is more plausible. For some parents, time spent with the child may be experienced the way it is by many grandparents: it is good in small doses but soon becomes exhausting. At the same time, these parents may not worry much about the risk of not spending any time at all with the child (risk neutrality). Other parents might differ in both respects, generating the same cardinal utility

function. To repeat, or re-repeat, these statements cannot (so far) be made rigorous, but they make obvious sense.[7]

### Rational beliefs

This concludes the discussion of the first component of a rational choice: choosing the best means to realize one's desires, given one's beliefs. Clearly, this is only a necessary condition for rationality, not a sufficient one. If I want to kill my neighbor and believe the best way of killing someone is to make a puppet representing him and stick a pin through it, I act rationally (as far as this first component goes) if I make a puppet representing my neighbor and stick a pin through it. Barring special circumstances, however, that belief is hardly rational.[8]

Rational beliefs are those that are shaped by processing the available evidence using procedures that, in the long run and on average, are most likely to yield true beliefs. Suppose we want to form a belief about the likelihood of rain on November 29, one week from today. We can probably not do much better than look up the statistics of rainfall in earlier years and assume that the (expected) future will be like the past. But as November 29 approaches, current rainfall may make us modify our expectations. If it often rains in November and we experience day after day with unclouded skies, we might infer the existence of a high-pressure system that makes rain on November 29 somewhat less likely.

This process of belief revision is often called *Bayesian learning* (named after the eighteenth-century minister Bayes). Assume that we have an initial ("prior") subjective probability distribution over different states of the world. In the example just given, the prior distribution was derived from past frequencies. In other cases, it might be a mere hunch. On the basis of my intuition, I might assign, for instance, probability 60 percent to the prime minister's (PM's) being competent and 40 percent to his being incompetent. We can then observe the actions he takes in office and their outcomes, such as the rate of growth of the economy. Suppose we can form an estimate about the likelihood of these observations *given the competence* of the PM. With a competent PM we have an 80 percent expectation of a good outcome, with an incompetent only 30 percent. Bayes showed how we can then update our initial probabilities concerning the PM's competence, *given the observations*.

[7] Hence the analogy with temperature scales is only partly valid. These scales measure *only* the intensity of temperature. Cardinal utility functions measure the joint result of intensity of preference and risk attitudes.

[8] Belief in witchcraft may be self-fulfilling, if the cursed person believes in the efficacy of the curse and simply loses the will to live. In that case, the observed efficacy of the curse might make belief in witchcraft rational, even if (as with the theory of action at a distance) the agent cannot specify the mechanism by which it works. It could also make witchcraft punishable on the basis of its actual consequences rather than, as suggested by Donne and Hobbes (see introduction to Part II), on the basis of *mens rea* only.

Assume that there are only two possible outcomes, good or bad, and that we observe a good one. If we write p(a) for the probability that a obtains and p(a | b) for the *conditional probability* that a obtains given that b obtains, we have assumed that p(PM is competent) = 60 percent, p(PM is incompetent) = 40 percent, p(good outcome | PM is competent) = 80 percent, and p(good outcome | PM is incompetent) = 30 percent. We seek to determine p(PM is competent | good outcome). We use the letters a and b to denote, respectively, competence and good outcome. We then note first that

$$p(a \mid b) = p(a \,\&\, b)/p(b) \qquad (*)$$

In words, the conditional probability p(a | b) equals the probability that both a and b obtain, divided by the probability of b. This follows from the more intuitive idea that p(a & b) equals p(b) multiplied by p(a | b). Dividing both sides of this equation by p(b), we get equation (*).

Using equation (*) again, but with a and b reversed, we have

$$p(b \mid a) = p(a \,\&\, b)/p(a)$$

or, equivalently,

$$p(a \,\&\, b) = p(b \mid a) \cdot p(a)$$

Substituting the latter expression in (*), we obtain

$$p(a \mid b) = p(a \mid b) \cdot p(a)/p(b) \qquad (**)$$

Now, there are two ways for b (the good outcome) to occur, with a competent PM or with an incompetent PM. Drawing on the fact that the probability that one of two mutually exclusive events will occur is the sum of the probabilities for each event, we can thus write

$$p(b) = p(b \,\&\, a) + p(b \,\&\, \text{not-a})$$

which, by the reasoning in the paragraph following (*), is equivalent to

$$= p(b \mid a) \cdot p(a) + p(b \mid \text{not-a}) \cdot p(\text{not-a})$$

If we substitute this expression for p(b) into (**), we obtain *Bayes's theorem*:

$$p(a \mid b) = p(b \mid a) \cdot p(a)/[p(b \mid a) \cdot p(a) + p(b \mid \text{not-a}) \cdot p(\text{not-a})]$$[9]

[9] If p(a) = 1, the formula yields p(a | b) = p(a), for any b. In other words, complete certainty is impermeable to new evidence. In particular, *fanatics* can never be persuaded that they are wrong. In Hume's example, "An English Whig, who asserts the reality of the popish plot, an Irish Catholic, who denies the massacre in 1641 and a Scotch Jacobite, who maintains the innocence of queen Mary, must be considered as men beyond the reach of argument or reason, and must be left to their prejudices."

Plugging in the numerical probabilities on the right-hand side of this equation tells us that p(a | b) = 80 percent, that is, that the observation of a successful outcome raises the likelihood that the PM is competent from 60 percent to 80 percent. A second and a third positive observation would raise it to 91 percent and then to 97 percent. If another person initially estimated p(a) = 0.3 rather than 0.6, three successive positive observations would raise her estimate first to 0.53, then to 0.75, and finally to 0.89. Hence it may not matter much whether the initial hunches are unreliable, since as more and more information comes in the updated beliefs become more and more trustworthy. Over time, initial differences of opinion can be swamped by new evidence.[10] For future reference (Chapter 22), we also note that each new piece of information has less of an impact than the previous one.

### Optimal investment in information-gathering

The third component of a rational action is the optimal investment of resources – such as time or money – in acquiring more information. As shown in Figure 13.1 there are several determinants of this optimum. First, how much information it is rational to acquire depends on the desires of the agent.[11] For instance, an agent who does not care much about rewards in the distant future would not invest much in determining the expected lifetime of a durable consumption good. More obviously, it makes sense to gather more information before making an important decision such as buying a house than when choosing between two equally expensive bottles of wine. In the latter case, one should perhaps just decide by flipping a coin, if the expected cost of determining which is the better exceeds the expected benefit (based on a prior knowledge of the quality range of equally priced wines) from drinking the better wine rather than the inferior one.

Desires and prior beliefs jointly determine the expected benefits from new information. It is sometimes possible to tell with great precision how many

[10] For the convergence to occur, the successive pieces of new information must be statistically independent of each other. In the textbook example of Bayesian belief formation, a person draws balls from an urn known to be equally likely to contain either 80 percent black and 20 percent white balls or 20 percent black and 80 percent white in order to determine how likely it is that the one or the other obtains. Since the draws are random and the balls are put back into the urn after each draw, the outcome of each draw is independent of the previous ones. In political situations such as the one described in the text, it may be much more difficult to verify independence. Also, convergence presupposes that the underlying situation remains the same or at least does not change too fast. If the environment changes rapidly, the process of updating beliefs resembles that of aiming at a moving target (Chapter 6).

[11] Wishful thinking, in which "the wish is the father of the thought," is clearly irrational. By contrast, there is nothing irrational about the process shown in Figure 13.1, in which the desires are, as it were, the grandfather of the beliefs.

additional lives will be saved by doing a specific test for cancer or, translated to the level of the agent, how likely it is that his or her life will be saved. The value of life depends on how the agent goes about trading off life against other desired ends. By one calculation, a premium of about $200 per year was required to induce men in risky occupations such as coal mining to accept one chance in a thousand per year of accidental death. Hence at the time this calculation was done, the value of a life was about $200,000.[12] The expected costs of new information, which are determined by prior beliefs, can also sometimes be ascertained with precision. To detect intestinal cancer, it is common to perform a series of six inexpensive tests on a person's stool. The benefits of the first two tests are significant. However, for each of the last four tests the costs of detecting an additional case of cancer (not even curing it) were found to be $49,150, $469,534, $4,724,695, and $47,107,214, respectively.

The optimal search for information may also depend on the results of the search itself (this is represented by the loop in Figure 13.1). When a new medical product is being tested, there is a prior decision to provide the medication to one group and withhold it from another for a certain period. If it becomes evident early on, however, that the product is spectacularly successful, it would be unethical to withhold it from the control group. The same argument applies to a single rational agent. Suppose I am out in the woods plucking berries. I know that berries tend to grow in clusters, so I am prepared to spend some time looking before I start plucking. If I am lucky and find an abundant patch right at the beginning, I would be foolish to keep on looking.

We may view the gathering of information as a *shadow action* that accompanies the primary action. Before deciding what to do, we have to decide how much information to collect. Sometimes, *the shadow action and the primary action may coincide*, at least partially. Suppose the leaders of a country are weighing whether to go to war against another country. Germany's invasion of France in 1940 can serve as an example. To make the final decision whether to attack, information was crucial. The leaders needed to know the objective capacities of the prospective enemy, as well as "the organization, customs and habits of the enemy's army" (from the German manual *Duties of the General Staff*). Much of this information could be gathered by conventional means, including spying. However, to determine the *morale* of the enemy – their fighting spirit – there was no other option than actually fighting them.

[12] There are many pitfalls in making such calculations, but the general point is impossible to deny: we all attach a finite value to our lives. If we did not, we would not engage in all the enjoyable or profitable risky activities that we do.

### Indeterminacy

These last examples – plucking berries and planning for war – will also help us see the *limitations* of rational-choice theory, or rather one of its two limitations. As an explanatory tool, the theory can fail in one of two ways. On the one hand, it may fail to yield unique predictions about what, in a given situation, people will do. On the other hand, people may fail to live up to its predictions, whether unique or not. The second failure, *irrationality*, is the topic of the next chapter. The first, *indeterminacy*, is the topic of the following remarks.

An agent may be unable to identify the best element in the feasible set, for one of two reasons. A consumer may be *indifferent* between two options that are equally and maximally good. In trivial cases, this happens when the options are indistinguishable, as when a consumer faces the choice between two identical cans of soup in the supermarket. In non-trivial cases, two options might differ along several dimensions so that the differences exactly offset each other. The non-trivial case is rare, perhaps non-existent. If offered a choice between two cars that differ in price, comfort, appearance, speed, and so on, I might not prefer either to the other without, for that matter, being indifferent between them. If I were, a five-cent discount on one car should induce a preference for that option. Intuition suggests that this is unlikely to happen.

In fact, the consumer's preferences may be *incomplete*. Suppose I have inspected five car models, A, B, C, D, and E, and rank them as shown in Figure 13.4 (arrows standing for the preference relation). My inability to compare C and D does not matter, since I am not going to buy either of them anyway. By contrast, my inability to compare A and B leaves me in a pickle. True, I might try to gather more information, but how do I know it is worth the trouble? I return to this point shortly.

First, however, let me point to another and probably more important source of incomplete preferences. Typically, option preferences are induced by outcome preferences. I prefer one option because I prefer its outcome, that is, its expected utility, compared to that of other options. If the situation is one of

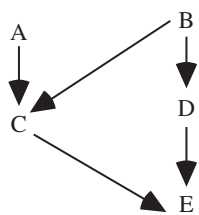



Figure 13.4

uncertainty or ignorance rather than risk (Chapter 7), however, I may not be able to compare the outcomes.[13] In the immortal words of Dr. Johnson, "Life is not long, and too much of it must not pass in idle deliberation how it shall be spent: deliberation, which those who begin it by prudence, and continue it with subtlety, must, after long expence of thought, conclude by chance. To prefer one future mode of life to another, upon just reasons, requires faculties which it has not pleased our Creator to give to us."[14] In the equally memorable words of Keynes, "Most, probably, of our decisions to do something positive, the full consequences of which will be drawn out over many days to come, can only be taken as a result of animal spirits – of a spontaneous urge to action rather than inaction, and not as the outcome of a weighted average of quantitative benefits multiplied by quantitative probabilities."

A further indeterminacy arises in the difficulty of determining the optimal investment in information gathering. When I am out plucking berries in unknown territory, how long should I keep looking for a dense cluster and when should I begin plucking? Unless I find a rich patch right away, it makes sense to spend some time looking around. At the same time, I do not want to keep looking until nightfall, because then I shall certainly go home with an empty basket. Between the lower and the upper bounds on the time one should spend looking, there may be a large interval of indeterminacy. A different problem arises in the case of planning for war. If the primary decision and the shadow decision coincide, the planner is doomed to remain in a state of (at least partial) uncertainty. Rational-choice theory cannot guide us well in these situations. The theory is helpful in highly structured situations about which a great deal is known, such as testing for cancer, but less so in unknown environments.

Whereas spending less time than the lower bound and more than the upper bound would be irrational, no choice the agent makes *within* this interval can be characterized as irrational. One might therefore, perhaps, think about dropping the idea of rationality in favor of that of *non-irrationality*. This revised version of the theory would allow us to make sense of a greater range of behavior but have less predictive power. Most practitioners of the theory

[13] In decision making under uncertainty, I may be able to compare options if the worst outcome of one option is better than the best outcome of another. In decision making under ignorance, even this modest comparability is unavailable.

[14] I might, however, "conclude by chance" and then invent the "just reasons," for instance, by giving greater weight to the attributes on which the chosen option is clearly superior. This can have undesirable consequences. Suppose I have the choice between going to law school and going to forestry school. Being unable to make a reason-based choice, I go to law school more or less by chance and justify the decision retrospectively by giving more weight to the income dimension of the two careers. With these newly induced preferences, I might go on to make other decisions that differ from those I would have made on the basis of my pre-choice preferences.

would, I believe, be reluctant to revise the theory along these lines. What attracts them to the theory in the first place is precisely that it holds out the promise of generating *unique* predictions. It does so by virtue of the elementary mathematical fact that any "well-behaved" utility function defined over a "well-behaved" opportunity set attains a maximal value for a unique member of that set. The interaction between opportunity set and indifference curves in Figure 10.1 offers a good example of the compelling simplicity of this idea, "doing as well as one can."

Indeterminacy can also arise if the feasible set is unknown and unknowable, as is the case in the search for innovations. The "innovation-possibility frontier" that some economists stipulated to propose a rational-choice explanation of technical change is an essentially meaningless concept. More generally, *creativity* cannot be reduced to a rational procedure. This proposition is obvious when it comes to artistic creation (Chapter 16), but applies to other cases as well. Scholars have proposed the idea of *integrative bargaining*, to move from a zero-sum activity in which one party's loss is the other party's gain to one in which both parties profit. An often-cited example is that of two sisters who were bargaining over an orange and split it in half. One sister who wanted only the juice, squeezed it out, drank it, and threw out the peel. The other sister, who wanted only the peel for a cake she was baking, threw out the pulp. Clearly, both could have done better. Yet in a given case, a mutually beneficial solution of this kind may not exist. A clever arbitrator may be able to think of one – or not.

Finally, an important source of belief indeterminacy arises in strategic interaction, when each agent has to form beliefs about what others are likely to do on the basis of *their* beliefs, knowing that they are going through similar reasoning with regard to his. In some cases, further considered in Chapter 18, the reward structure does not allow the agents to converge to a commonly held set of beliefs.

### Rational choice under uncertainty

Often, agents are not able to make a rational choice in conditions of uncertainty. In some cases, though, rationality under uncertainty is well defined. Thus in England in 1797–8, the fear of an imminent French invasion bought the value of consols (government bonds) down to half their price. A smart investor, Thomas Thompson of Hull, decided that "if the French landed it mattered not whether he met his fate as a rich man or a poor one," and invested heavily in the funds. When the invasion threat vanished and consols revived, he made a killing. The situation may perhaps be reconstructed as in Table 13.1.

In the language of economists, the option of buying the consols has *stochastic dominance*: it can never be worse than not buying, and it can be better.

Table 13.1

| | The French invade | The French do not invade |
|---|---|---|
| Buy consols | Die poor | Large gain |
| Not buy consols | Die rich | No gain |

!

Since this statement makes no appeal to probabilities, it is consistent with the investor being in a state of complete uncertainty about the invasion. In experimental treatments of fatal diseases with no known cure, it is also rational to apply the principle "Can't hurt, might help" without any knowledge about the prospects of a successful outcome. A murderer who would get the death penalty if caught might rationally decide to kill any witnesses to his act, reasoning that he might as well be hanged for a sheep as for a lamb. This decision, too, requires no probabilistic reasoning, except for the tacit assumption that a multiple murder will not trigger a more intensive search with a higher probability of capture. If that assumption is unjustified, rational-choice theory will not tell the murderer what to do. He cannot compare the expected values of committing one murder and committing several, since he only knows the *ordinal* probabilities (Chapter 7) of being caught if he makes the one and the other choice. Calculation of expected value requires both cardinal utilities and cardinal probabilities.

### Rationality is subjective through and through

Let me conclude the discussion of rational-choice theory by emphasizing again its *radically subjective nature.* One might, to be sure, take the word "rational" in an objective sense, implying that a rational agent is one who makes decisions that *make his life go better* as judged by objective criteria such as health, longevity, or income. Used in this way, however, the idea would not have any explanatory power. As I have emphasized, *consequences* of a decision cannot explain it. Only the mental states that precede the decision enable us to *explain* the actions as optimal from the point of view of the agent rather than to *characterize* them as useful or beneficial from the point of view of an external observer (or of the agent at a later time).

Suppose I suffer from a severe inability to defer gratification, that is, from being unable to take account of future consequences of present behavior. And suppose scientists came up with a discounting pill, which would increase the weight of future rewards in present decisions. If I take the pill, my life will go better. My parents will be happy I took the pill. In retrospect, I will be grateful that I did. But if I have a choice to take the pill or not, I will refuse if I am rational. Any behavior that the pill would induce is already within my reach. I could stop smoking, start exercising, or start saving right now, but I do not. Since I do not want to do it, I would not want to take a pill that made me do it. Similarly, a selfish person would refuse an "altruism pill" and, even more compellingly, an altruistic person a "selfishness pill." If I love my family and am willing to sacrifice some of my hedonic welfare for their sake, I would refuse a pill with the two-step effect of lowering theirs, just as I would refuse any option (e.g. buying an expensive meal for myself) that produced the same effect in one step.

To sharpen the argument, assume that a person consumes x today and y tomorrow, and that her one-period discount rate is 0.5 (she is indifferent between one unit of utility tomorrow and one half-unit today). Assume for simplicity that $u(x) = x$ and $u(y) = y$. The discounted presented value of her consumption stream is $x + 0.5y$. Suppose the person learns that tomorrow she is going to suffer from pain that will reduce the utility of her consumption by a factor of 0.5. The discounted present value of consumption now is $x + 0.25y$. If a rational agent is offered a costless aspirin that will eliminate the pain, she would clearly take it, thus restoring the original present value. If she took a pill that induced a discounting rate of 1 (but did not take the aspirin) the outcome would be the same in the sense that in either case, she would be indifferent between the two-period stream and a period-one utility of $x + 0.5y$. Since the agent would take the aspirin and since its effect is the same as that of the discounting pill, why would she not take the latter? The reason is that the choice of the discount pill is constrained by the need for the pill-induced consumption to be superior to non-pill consumption *as judged by prepill preferences*. There is no similar constraint on the aspirin choice, because there is no difference between pre-aspirin and post-aspirin preferences. Even without the aspirin I prefer being able to be free of pain tomorrow. When that state becomes part of my repertoire, I choose to bring it about. By contrast, the utility stream induced by the discounting pill is already in my repertoire, but I choose not to bring it about.[15]

[15] With hyperbolic discounting, an agent might accept a discounting pill. Using the numerical example from Chapter 6, suppose that the effect of the pill is to lower the value of k from 1 to 0.3. At the time the smaller reward becomes available, its present value is simply 10 (no discounting). The present value of the larger reward of that time is $30/(1 + 0.3.5) = 12$. Hence

Choices, in other words, need to be seen through the eyes of the agent. A myopic person who loses his glasses may be prevented by his myopia from finding them. He is "trapped."[16] Similarly, a rational agent may find himself in a "belief trap" that leaves him stuck with a false belief, namely, if the believed costs of testing the belief are too high. Thus women who practice genital mutilation may be caught in a belief trap. The Bambara of Mali believe that the clitoris will kill a man if it has contact with the penis during intercourse. In Nigeria, some groups believe that if a baby's head touches the clitoris during delivery, the baby will die. In Poland it has been widely believed that anyone who drinks when using disulfiram (Antabuse) implanted under the skin will die. As a matter of fact, implanted disulfiram is pharmacologically inert. The false belief might nevertheless deter people from testing it.

The *rationality* of beliefs is a completely different matter from that of their *truth*. Whereas truth is a feature of the relation between the belief and the world, rationality is a feature of the relation between the belief and the evidence possessed by the agent. Although rationality may require the agent to invest in new information, the investment is always constrained by its expected (that is, *believed*) costs and benefits. If gathering more information is believed to have high *opportunity* costs, as in the face of a possible imminent danger, it may be rational to abstain from the investment. If it is believed to have high *direct* costs, as in testing the belief about the fatal effects of drinking while using implanted disulfiram, only an irrational person would make the investment. More generally, many beliefs must be taken at face value secondhand, since if we were to test them all we would never get on with our lives.

Any choice-based explanation of behavior is subjective. Not all subjective explanations assume, however, the transparency of the agents to themselves and the relentless search for optimality that are the hallmarks of rational-choice explanations. In the next chapter I shall canvass a number of explanations that depart from rational-choice theory on one or both accounts.

precommitment in the form of taking the pill will enable the agent to act in accordance with his calm and reflective judgment, thus preventing weakness of will (in the broad sense). This statement remains true even if he has to buy the pill, as long as its cost (in utility terms) is less than 2. It also remains true if precommitment has the effect of reducing the value of the delayed reward (perhaps the discounting pill has the side effect of reducing the capacity for enjoyment), as long as the loss is less than 5. These facts might be relevant if for the discounting pill we substitute psychotherapy.

[16] If offered his glasses, he would put them on. I have argued that if offered the discounting pill, he would not take it. The difference is that he can already do without the discounting pill anything he could do if he took it, whereas there are many things he cannot do without his glasses that he could do if he put them on.

### Bibliographical note

I discuss the relation between reason (in the sense of Chapter 4) and rationality in my inaugural lecture at the Collège de France, *Raison et raisons* (Paris: Fayard, 2006). For more about Weber and rationality, see my "Rationality, economy, and society," in S. Turner (ed.), *The Cambridge Companion to Weber* (Cambridge University Press, 2000). A classic exposition of utility theory is found in R. D. Luce and H. Raiffa, *Games and Decisions* (New York: Wiley, 1957). The original work by J. von Neumann and O. Morgenstern, *The Theory of Games and Economic Behavior*, 2nd edn (Princeton University Press, 1947), is still worth consulting. An outstanding exposition of rational-choice theory (and its problems) is R. Hastie and R. Dawes, *Rational Choice in an Uncertain World* (Thousand Oaks, CA: SAGE, 2001). I discuss the child custody example at greater length in Chapter 3 of *Solomonic Judgments* (Cambridge University Press, 1989). An excellent elementary presentation of Bayesian theory is R. Winkler, *An Introduction to Bayesian Inference and Decision* (Gainesville, FL: Probabilistic Publishing, 2003). I discuss and criticize the idea of an "innovation possibility frontier" in *Explaining Technical Change* (Cambridge University Press, 1983), pp. 104–5. The story about Thomas Thompson is taken from J. Uglow, *In These Times* (London: Faber and Faber, 2014), p. 223. My argument that a rational person would not take the discounting pill has been influenced by exchanges with Gary Becker and Peter Diamond; see also O. J. Skog, "Theorizing about patience formation: the necessity of conceptual distinctions," *Economics and Philosophy* 17 (2001), 207–19. I take the idea of a belief trap from G. Mackie, "Ending footbinding and infibulation: a convention account," *American Sociological Review* 61 (1996), 999–1017. A useful study of the importance of intelligence in preparing for war is E. R. May, *Strange Victory: Hitler's Conquest of France* (New York: Hill & Wang, 2000). I owe the information about the use of implanted disulfiram in Poland to W. Osiatynski, *Alcoholism: Sin or Disease?* (Warsaw: Stefan Batory Foundation, 1997), and the data about its ineffectiveness to J. Johnsen and J. Mørland, "Depot preparations of disulfiram: experimental and clinical results," *Acta Psychiatrica Scandinavica* 86 (1992), 27–30.

# 14 Rationality and behavior

### Ignoring the costs of decision making

The idea of rationality has a strong normative appeal. We *want* to have reasons – desires and beliefs in light of which the action appears as rational – for what we do. In fact, our desire to act for a reason – our deference to rationality – can be so strong as to induce irrational behavior. We may define *hyperrationality* as the propensity to search for the abstractly optimal decision, that is, the decision that would be optimal if we were to ignore the costs of the decision-making process itself. These costs are of three kinds: (1) the cost of the means of deciding, (2) the cost of the side effects of deciding, and (3) the opportunity costs of deciding, that is, the value of the other things one might have done instead of going through the decision process. Let me illustrate them briefly.

Hyperrationality through neglect of (1) could arise in *comparison shopping* when the (expected) savings from finding the lowest price is less than the money spent on transportation traveling from store to store. Tourists in the south of France cross the border to Spain to buy cheap cigarettes as if gasoline were free.[1] Neglect of (2) could induce hyperrationality in contested child custody cases. The court may try to promote the interest of the child by determining which parent is more fit for custody.[2] Once that issue has been settled, the court has a good reason for awarding custody to that parent. In the juridico-psychological process of ascertaining relative fitness, however, incalculable damage may be done to the child. A more rational procedure, given the aim to be achieved, might be to flip a coin or retain the traditional presumption of maternal custody.

[1] At a restaurant in Baltimore, the waitress announced that over the next five minutes all drinks would be sold at half price. When I asked about the price of the most expensive item, she indicated Johnny Walker Blue whisky, at $35 dollars a glass. I then asked her whether some clients had ever ordered this drink, even if they did not like whisky. She answered in the affirmative. These clients acted, that is, to maximize savings rather than utility. Their reaction can be compared to that of the subjects in an experiment discussed later, who acted to minimize waste rather than to maximize utility.

[2] The "best interest of the child" is in fact the criterion used in most child custody laws.

Neglect of opportunity costs is illustrated in an observation by Dr. Johnson in a conversation with Boswell about which subjects children should be taught first: "Sir, it is no matter what you teach them first, any more than what leg you shall put into your breeches first. Sir, you may stand disputing which it is best to put in first, but in the mean time your breech is bare. Sir, while you are considering which of two things you should teach your child first, another boy has learnt them both."[3] Again, flipping a coin may be more rational. Or consider the doctor who arrives at the scene of an accident and has to decide what steps to take. Although he obviously needs to examine the patient, his behavior is self-defeating if he spends so much time on it that the patient dies under his hand. When confronted with a patient suffering from a chemical injury to the eye, "the ophthalmologist should not appreciate history-taking as his initial goal; instead, he must 'shoot first and ask later': details of injury are asked during or after the initial irrigation." Others may have had the experience, when plucking berries, of looking so long for the best place that by the time they find it night is falling. Even when the savings from comparison shopping exceed the transportation costs, the behavior might still be irrational because of the value of the lost time.

### Some canonical principles of rationality

By a "puzzle" I shall here understand observed behavior that seems recalcitrant to rational-choice explanation. Although some puzzles may, on closer inspection, lose their puzzling character, many do not. Experiments and real-life behavior show numerous instances of behavioral patterns that violate the canons of rationality. In the following selective list of these canonical principles, I begin with the more fundamental and proceed to the more specific. I limit myself to individual choices; anomalies in interactive choices are discussed in Part IV. The list may be supplemented by some of the cognitive anomalies I discussed in Chapter 7.

1. In a choice between acting and doing nothing, a rational agent will not act if the expected utility costs of acting exceed the expected utility benefits.
2. In the choice between evils, a rational agent will choose the lesser evil.
3. A rational agent assigns the same weight to opportunity costs and to direct costs.
4. A rational agent will never prefer having a subset of a set of options to having the full set.

[3] Johnson, being a Shakespeare scholar, may have had in mind the king's remark in *Hamlet*, Act 3, scene 2: "Like a man to double business bound I stand in pause where I shall first begin, and both neglect."

5. If a rational agent prefers X to a glass described as half-full she should also prefer X to one described as half-empty.
6. In a game of pure chance, a rational gambler will not, when placing her bets, pay attention to the outcomes of previous gambles.
7. When deciding whether to persist in a project or scrap it, a rational investor will pay attention only to the present value of future utility streams from these two options.
8. If at time 1 a rational agent forms a plan to carry out action X at time 2, she will do X at time 2 unless either her desires or her beliefs have changed in the meantime.
9. In a risky choice, a rational agent will choose means according to the expected outcome, not only according to the best-case (or worst-case) scenario.
10. In a market of rational agents, the rate of return on all assets should be (approximately) the same.[4]
11. If a rational agent chooses A from the set (A, B, C), she will also choose A from the set (A, B).
12. A rational agent will not act on an effect to suppress the cause (she will take antibiotics rather than aspirin to cure pneumonia).
13. If a rational agent prefers getting reward X with certainty to getting reward Y with probability q, she will also prefer getting X with probability p to getting Y with probability pq (the independence axiom of cardinal utility theory).
14. If a rational agent does X when she knows that circumstance C obtains (or intends to do X when C is expected to obtain) and does X when circumstance C does not obtain (or intends to do X when C is not expected to obtain), she should do or intend to do X even when she is ignorant about the circumstances.
15. A rational agent will never make an offer if its acceptance will reveal information that makes the deal have negative expected value.
16. If an offense induces a desire for vengeance, the offended person will, if rational, bide his time until he can strike back with maximal chance of success or with minimal risk for himself.[5]
17. If challenged to a fencing duel, a rational agent will take fencing lessons if he has to take up the challenge.

[4] There are two reasons why the equality can be expected to be only approximate. First, external shocks will always induce deviations from equality. Second, risk aversion may induce lower values (and therefore higher rates of return) on highly volatile assets.

[5] Some might be tempted to replace "or" with "and" in this sentence. Except by fluke, however, one cannot maximize two objectives at the same time. To be more precise, the agent would seek the optimal feasible mix of the two goals, as represented by an opportunity set and a family of indifference curves (Chapter 10).

18. Before asking for another person's hand in marriage, a rational agent will gather information about the other's behavioral and emotional propensities.
19. In updating beliefs, a rational agent should reach the same conclusion when receiving one piece of information before another as when receiving them in the opposite order.

### Violations of the canon

These normatively compelling principles are, it turns out, routinely violated. Examples (with numbers matching those of the principles they violate) follow.[6]

1. *The paradox of voting*. Since no national election has ever been won by a single vote, an individual vote makes no difference to the outcome and may entail considerable trouble for the voter. Yet people do vote in large numbers.[7]
2. *More pain preferred to less*. As noted in Chapter 6, subjects who were exposed to two sequences of highly unpleasant noise chose, when asked which they would prefer to be repeated, the one that was unambiguously less pleasant.
3. *The lawn-mowing paradox*. In a small suburban community, Mr. H. mows his own lawn. His neighbor's son would mow it for $12. He would not mow his neighbor's same-sized lawn for $20.
4. *The Christmas club puzzle*. In this system, customers deposit a monthly sum at low or no interest, which they can only withdraw at Christmas. The option of earning normal interest and costless withdrawal at will is also open to them.
5. *The credit card paradox*. When credit cards were introduced, the credit card lobby preferred ~~that~~ any difference between cash and credit card customers be labeled as a cash discount rather than as a credit card surcharge. Although the two descriptions are logically equivalent, consumers were more likely to use the cards if the difference was described as a cash discount.

[6] Many of the examples have been cited in previous chapters and summarized here for convenience. They are from various sources: proverbs, classical authors, thought experiments, laboratory experiments, and real-life observations. Examples in the first three categories are, however, based on well-established theories that are surveyed later in the chapter.

[7] The paradox arises when the sole aim of the voters is to put a candidate into office or a proposal into effect. It need not arise when the aim is to contribute to the vitality of the democratic system or to give a "mandate" to a candidate, since in these cases votes matter even if they are not pivotal. Yet even when the motivation is to support democracy the explanation may lie in one of the non-rational mechanisms discussed later.

6. *Two gamblers' fallacies*. If red has come up five times in a row, about one-half of gamblers believe that it is more than 50 percent likely to come up black next time. The other half believes it is less than 50 percent likely to come up black.
7. *The sunk-cost fallacy*. If you buy tickets for an event and heavy snowfall makes it burdensome to get there, you might still go even though you would have refused the tickets had they been offered to you free.

   The fallacy is illustrated by an experiment in which the subjects were asked the following question:

   > Assume that you have spent $100 on a ticket for a weekend ski trip to Michigan. Several weeks later you buy a $50 ticket for a weekend ski trip to Wisconsin. You think you will enjoy the Wisconsin ski trip more than the Michigan ski trip. As you are putting your just-purchased Wisconsin ski trip ticket in your wallet, you notice that the Michigan ski trip and the Wisconsin ski trip are for the same weekend! It's too late to sell either ticket, and you cannot return either one. You must use one ticket and not the other. Which ski trip will you go on?

   Thirty-three subjects preferred the trip to Michigan, twenty-eight the trip to Wisconsin. According to rational-choice theory, however, *all* should have chosen to go to Wisconsin. Thus over half of them acted as if they wanted to minimize waste rather than maximize utility.
8. *The dentist puzzle*. On March 1 I make an appointment with the dentist for April 1. On March 30 I call her to say that because of a (fictitious) funeral in the family I cannot keep it. Except for the sheer passage of time, no change has occurred in the interval. In particular, the pain from toothache is the same.
9. *Best- and worst-case scenarios*. Cancer patients in late stages often overestimate their chance of survival. Rather than palliative therapy to relieve their pain, they choose aggressive and painful chemotherapy with few benefits. When asked how much they would pay to reduce the likelihood of a low-probability disaster, people are willing to pay as much to have it reduced to one chance in 1 million as to have it reduced to one chance in 10 million.
10. *The equity premium puzzle*. Historically, the yield on stocks is vastly higher than the yield on bonds. A person who invested $1 in stocks on January 1, 1928, would on January 1, 1998, have a portfolio worth $1,800. Somebody who invested $1 in bonds would have a portfolio worth $15. The puzzle is why this discrepancy has not led to a rise in the value of stocks to bring the return on stocks closer to the return on bonds.
11. *Effect of irrelevant alternatives*. If each of two options A and B is superior to the other along one of two relevant dimensions, people may find it hard to choose and instead decide to gather more information about the options.

If a third option C, which is (1) inferior to A along both dimensions and (2) inferior to B on one dimension and superior on another, is introduced, there is a tendency to choose A without further search.

12. *The cold-water puzzle* (Chapter 7). In an experiment, subjects who were led to believe that the length of time they could hold their arms in painfully cold water was the best indicator of longevity held their arms in the water longer than those not given this (false) information.
13. *The certainty effect* (Chapter 7). In experiments, a majority prefer to win a one-week tour of England with certainty to a 50 percent chance of winning a three-week tour of England, France, and Italy, but a majority also prefer a 5 percent chance of the second option to a 10 percent chance of the first.
14. *The disjunction effect*. If subjects in an experiment expect to win in a gamble and are asked whether they will agree to take part in a further gamble, they tend to say yes. If they expect to lose, they are likely to state the same intention. If they do not know whether they will win or lose, they are less likely to do so. The same effect is observed in one-shot Prisoner's Dilemmas: a person is more likely to cooperate if he knows that the other cooperated than if he knows he defected, and – this is the disjunction effect – *even more likely* to cooperate if he is ignorant of the other's choice.[8]
15. *The Winner's Curse*. In this experiment, subjects are asked to bid for a piece of land and told that the seller knows its exact value, whereas they know only that the value falls within a certain range, with all numerical values in that range equally likely. Buyers are also told that if they acquire the piece of land, it will be worth 50 percent more to them than to the seller, because they will be able to exploit it more efficiently. If an offer is accepted, rational buyers should be able to infer *from that fact* that the expected value to them of the land is less than what they bid. If the values range from 0 to 1,000 and a bid of (say) 600 is accepted, the buyer can infer that the real value to the seller is between 0 and 600, with an expected value of 300. Hence its expected value to the buyer would be 450, which is less than what he offered to pay. Since the same argument can be made for any bid that is accepted, rational buyers should never make a bid. Yet in experiments (which were inspired by real cases) nobody fails to make a bid.
16. *Rush to vengeance*. A proverb has it that "vengeance is a dish that is best served cold." Another says that "delay in vengeance gives a heavier

[8] Emotions, too, can be subject to his effect. When people believe that an airplane explosion must have been due either to a terrorist act or to poor maintenance, their anger will be triggered only when their belief is fixed on one of the explanations.

blow." Presumably, both arose in reaction against vengeance in hot blood, thus testifying to the existence of that phenomenon.

17. *Disregard for efficiency*. Montaigne wrote that "the honor of combat consists in rivalry of heart not of expertise; that is why I have seen some of my friends who are past masters in that exercise choosing for their duels weapons which deprived them of the means of exploiting their advantage and which depend entirely on fortune and steadfastness, so that nobody could attribute their victory to their fencing rather than to their valor."
18. *Marry in haste, repent at leisure*. This dictum applies not only to marriage in the literal sense. When people fall in love with a house, they are sometimes so eager to sign the contract that they fail to discover hidden flaws that surface later. In France, they have a week to change their mind; in Norway, the decision is irreversible.
19. *Order effects*. Contrary to Bayesian principles, belief updating is not neutral with respect to the order in which information is received, since early information may be either overvalued or undervalued (primacy and recency effects). Also, if strong evidence for guilt is followed by weak evidence for guilt, the belief of jurors moves *away* from guilt, contrary to Bayesian theory.

### Alternatives to rational-choice theory

To account for these puzzles, there is now available a wide repertoire of alternatives to rational-choice explanation. Before discussing them individually, let me list the key mechanisms in the alternative accounts I shall consider (with the puzzle numbers in parentheses). Some puzzles are listed more than once, because they may plausibly be accounted for in more than one way.

- Loss aversion (3, 5, 7, 10)
- Non-probabilistic weighting of outcomes (13)
- Hyperbolic discounting (4, 8)
- Biases and heuristics (2, 6, 19)
- Wishful thinking (9, 12)
- Inability to project (15)
- The desire to act for a reason (11, 14)
- Magical thinking (1, 12, 14)
- The categorical imperative (1)
- Emotions (3, 7, 14, 18)
- Social norms (1, 3, 16, 17)

In current thinking, the most prominent mechanisms are probably loss aversion and hyperbolic discounting. The first undermines the idea of *expected utility*; the second that of *discounted utility*. In my opinion, emotions are an even more

important source of irrational behavior, whether they operate directly or through the intermediary of social norms. Although emotions can upset rationality in many ways, the most important is perhaps by affecting belief formation, through wishful thinking, the hot–cold empathy gap, and urgency.

**Loss aversion** is defined with respect to a reference point, on the assumption that people attach value to changes from a given baseline rather than to the end states obtaining after the change. The reference point is typically taken to be the status quo, although subjects may be induced to choose other reference points. Loss aversion is the tendency for people to attach larger value (in absolute terms) to a loss from the reference level than to a same-sized gain.[9] Empirically, the ratio is found to be about 2.5 to 1, which I assume in the following. Another important property of the value function is that it is concave for gains and convex for losses, meaning that each extra unit of gain is valued less than the previous one, and each extra unit of loss is less painful than the previous one.

Two of the puzzles can be explained by the simple fact that losses loom larger than gains. To resolve the lawn-mowing puzzle, we need only observe that loss aversion predicts that opportunity costs and out-of-pocket expenses are valued very differently. Since the value to the homeowner of a gain of $20 is equivalent to the value of a loss of $8, he prefers forgoing the gain to paying $12 out of his pocket. The same reasoning could explain the credit card puzzle.

The resolution of the equity premium puzzle requires an additional premise, namely, that people choose their mix of bonds and stocks within a short time horizon. Since the return on stocks is volatile whereas bonds yield a steady income year in, year out, we can view the holding of stocks as accepting a risky gamble. Suppose we offer a person a bet on stocks that gives her a 50 percent chance to win $200 and a 50 percent chance to lose $100, with the fixed return to bonds as the reference point. If we assume loss aversion, this is captured by saying that the value of money is equal to x for $x > 0$ and equal to 2.5x for $x < 0$. Since the value of a loss of $100 is equal (in absolute terms) to a gain of $250, the prospect of a gain of $200 cannot compensate her for the equally likely prospect of a loss of $100. She will, therefore, reject the offer. Suppose now that we offer her a package of two such bets, to be carried out in successive periods. This compound gamble amounts to a 25 percent chance of gaining $400, a 50 percent chance of gaining $100, and a 25 percent chance of losing $200. If we multiply the loss by 2.5 to make it comparable to the gains

[9] Assuming that goods (including money) have decreasing marginal utility, standard rational-choice theory also predicts that losses will count more heavily than equal-sized gains from the same baseline. The magnitude of the effect, however, is typically much smaller. Also, standard utility theory implies that the utility gain of moving from A to B equals the utility loss of moving from B to A, since these differences are simply derived by comparing the utility *levels* of the two states.

and calculate the expected value,[10] it is easily shown to be 25. The person will, therefore, accept the compound gamble. Empirical studies suggest that investors do tend to reevaluate their portfolios too frequently, a myopic practice that induces them to invest too little in stocks and too much in bonds.[11]

A possible resolution of the sunk-cost puzzle appeals only to the curvature (convexity or concavity) of the value function. Let us consider the following example. A family pays \$p for tickets to a game to be played sixty miles way. On the day of the game there is a snowstorm. They decide to go anyway but note in passing that had the tickets been given to them, they might have stayed home. Writing v for the value function for gains, the value of going to the game is v(g). Writing v* for the value function for losses, the value of losing \$p is the negative number v*(–p). The cost of enduring the snowstorm is c. We assume that $v(g) = -v^*(-c)$, implying that if the family had received the tickets free they would have been indifferent between staying home and going to the game in a snowstorm. But since they have already paid \$p, they prefer to go. To see this, note first that because of the convexity of v*, $v^*(-(c + p)) > v^*(-c) + v^*(-p)$.[12] This can be rewritten as $v^*(-(c + p)) - v^*(-c) > v^*(-p)$, which on the assumption just stated is equivalent to $v^*(-(c + p)) + v(g) > v^*(-p)$. Since the left-hand term in the last inequality is the net gain or loss from going to the game and the right-hand term the loss from not going, they prefer to go.

Loss aversion may explain behavior, but since it is not very intuitive, might it not itself be in need of an explanation? The discoverers of loss aversion explain it by stating that "pain is more urgent than pleasure." In *The Theory of Moral Sentiment*, Adam Smith wrote that "Pain ... is, in almost all cases, a more pungent sensation than the opposite and correspondent [?] pleasure. The one, almost always, depresses us much more below the ordinary, or what may be called the natural state of our happiness, than the other ever raises us above it." But this cannot be the whole story, as some pleasures are surely more urgent and pungent than some pains. In his *Lectures on Jurisprudence,* Smith proposed a different account of the asymmetry: "It is a common saying, that he

[10] Since prospect theory assumes that the decision weights differ from probabilities, this calculation is only approximately correct. Neither this simplification nor the assumption of a linear value function matters for the conclusion of the analysis.

[11] The term "myopic" does not have the same meaning here as it has in analyses of time discounting (Chapter 6). It does not refer to the way the agent calculates the present value of future streams of income, but to the tendency to make successive decisions separately rather than "bundling" them together in one overall choice. Such "decision myopia," as we might call it, could also operate in other contexts. Thus when people try to control hyperbolic discounting by "bunching" successive choices together (Chapter 15), their success may depend on the number of choices they include.

[12] Since this is a comparison of two negative numbers, the inequality states that the former is closer to zero, that is, smaller in absolute terms.

who does not pay me what he owes me, does me as great an injury as he who takes as much from me by theft or robbery. It is very true the loss is as great, but we do not naturally look upon the injury as at all so heinous. One never has so great dependence on what is at the mercy or depends on the good faith of another as what depends only on his own skill." Whereas out-of-pocket expenses are certain, outside the laboratory gains foregone often have a more shadowy mental existence that may explain why they are less motivating.[13]

**Non-probabilistic weighting of outcomes**. Loss aversion follows from an influential alternative to rational-choice theory called *prospect theory*. Another implication of that theory is that people tend to weigh outcomes differently than expected utility theory asserts. According to that theory, utility is linear in probabilities (Chapter 12). Prospect theory, by contrast, argues that people are most sensitive to changes in probability near the natural boundaries of 0 (impossible) and 1 (certain). The certainty effect illustrates the non-linearity around 1. The creators of prospect theory, Daniel Kahneman and Amos Tversky, cite the following example (which they attribute to Richard Zeckhauser) of the non-linearity around 0:

> Suppose you are compelled to play Russian roulette, but are given the opportunity to purchase the removal of one bullet from the loaded gun. Would you pay as much to reduce the number of bullets from four to three as you would to reduce the number of bullets from one to zero? Most people feel that they would be willing to pay much more for a reduction of death from 1/6 to zero than for a reduction from 4/6 to 3/6. Economic considerations [that is, expected utility theory] would lead one to pay more in the latter case, where the value of money is presumably reduced by the considerable probability that one will not live to enjoy it.

**Hyperbolic discounting** was discussed in Chapter 6. Here, let me simply note the close link between puzzles (8) and (4). The reason people join Christmas clubs is presumably that they know that if they put their savings in a normal account with the intention of keeping them there until Christmas, they will change their mind and take them out again.

**Heuristics and biases**. Heuristics (rules of thumb) can lead people astray. The gambler's belief that the roulette wheel has a memory may stem either from the representativeness heuristic ("It is time for red to come up") or from the availability heuristic ("Red is on a roll"). The preference for the more unpleasant noise stems from the use of a "peak-end" heuristic.

[13] This difference might explain a seeming anomaly in the reactions of the Americans to British policy toward the American colonies in the eighteenth century. They accepted without any qualms Navigation Acts and many other policies that prevented the Americans from developing their trade and industry, such as the ban on export to America of utensils used in cotton and linen manufactures, while rising up in arms against what were (by comparison) insignificant attempts to raise revenue by taxation.

**Wishful thinking**. The phenomenon of wishful thinking was discussed in Chapter 7. It may be triggered by a simple desire, as when people in well-paid risky occupations downplay the risks they are running. It is even more likely to occur when the desire stems from a strong emotion, as when terminal cancer patients choose treatment whose only effect is to make them suffer more.

**Inability to project**. In a number of situations, people make bad decisions because of their inability to project themselves into the future. By this I mean the lack of ability to imagine what they or others would have reasons to believe, or incentives to do, in future situations that depend on their present choice. The Winner's Curse can be explained by this inability. For another example, consider President Chirac's disastrous calling of anticipated elections in June 1997. The reason his coalition lost may be that the voters understood that if he wanted early elections, it was because he knew something they did not and that made him believe that he would lose if he waited. By calling early elections, he revealed what he knew, or at least revealed *that* he knew something unfavorable, and therefore gave them a reason to vote against him. The polls told him he would win, but *polls are unlike elections* since holding a poll does not reveal anything to the respondents about the beliefs of the person who commissioned it.[14] The inability to project may also apply to the agent herself, if she makes a non-credible threat because she does not see that she would not have an incentive to carry it out if the target fails to comply.

**The desire to act for a reason**. I cited several instances of this mechanism in Chapter 9 and at the beginning of this chapter. In puzzle 11, the desire causes the agent to modify her behavior upon the introduction of an option that is unambiguously inferior to one of the options she already possesses.[15] In other cases, adding options may prevent the agent from making any decision at all. A psychologist who set up stalls on Broadway selling jam found that when stalls had a large variety of brands passers-by looked at more of them but purchased fewer, compared to stalls with few varieties. With more options it is more difficult to say to oneself, unhesitatingly, "This is the best." Those who need to base their choice on sufficient reasons will abstain from choosing.

As suggested by puzzle 14, to act for a reason one needs to *have* a reason, not merely to *know* that one has a reason. Thus suppose I know that exactly

[14] In Part IV I discuss Chirac's behavior as an instance of what I call the "younger sibling syndrome."

[15] This phrasing is somewhat misleading, since in experiments it is not the *same* subjects who are exposed first to the choice set (A, B) and then to set (A, B, C), but two different groups of subjects allocated at random to one of them. The natural interpretation of the finding, however, is that the subjects in the (A, B, C) group *would have behaved* as subjects in the (A, B) group had they been exposed only to the two options. The "modification," therefore, refers to a counterfactual baseline, not to an actual one. This remark applies to many of the experiments cited in this book.

one of p or q is the case, but I do not know which. If p is the case, I have a reason to do X. If q is the case, I also have a reason to do X. Hence I know that whatever is the case, I have a reason to do X, but since I do not know *which* reason, I abstain from X. Puzzle 14 offers an example of this anomaly.[16] For another example, consider the fact that in older English law an accused would be acquitted if the evidence left it uncertain whether he had committed theft or embezzlement, although he would have been convicted if either charge had been proven. To be convicted he would have to be found either guilty of p or guilty of q. Being found guilty of (p or q) would not be sufficient.

**Magical thinking**. The mechanism of magical thinking (Chapter 7) could explain behavior in the cold-water puzzle. It may also explain some cases of the disjunction effect. If people are more likely to cooperate in the Prisoner's Dilemma when they do not know whether the other person cooperated or defected, it may be because they believe, magically, that by cooperating they can bring about the cooperation of the other. "Being like me, he will act like me." Voting intentions, too, may be shaped in this way. If I believe, irrationally, that my voting is not merely a predictor of others' voting but somehow makes it more likely that they will vote, the increased efficacy of my action makes voting appear rational.

**The categorical imperative**. There is a close relation between this last instance of magical thinking and (an everyday version of) Kant's categorical imperative, according to which one should do A rather than B if we would all be better off if all did A than if all did B. Acting on the categorical imperative is, however, irrational. Rationality tells me to choose as a function of what will happen if *I* do A rather than B.[17] The categorical imperative tells me to choose as a function of what will happen if *everybody* does A rather than B. In a national election, even those who are not subject to magical thinking might "abstain from abstaining" by the thought "What if everybody did that?"

**Emotions**. To compare emotion-based behavior with rational behavior, we may modify Figure 13.1 to include (in the heavily drawn lines) the impact of emotion on each of the elements of the scheme (see Figure 14.1).

It has been argued that emotions may affect *action* directly, in cases of weakness of will (Chapter 6). Medea, when killing her children to take revenge on Jason, knows *as she is doing so* that she is acting against her better judgment. I noted my reservations about that idea, but it cannot be excluded. Emotions affect *desires* in two ways. First, by virtue of the associated action

[16] Along similar lines, the maximal amount people are willing to pay for a lottery between two options has been shown to be less than the maximum they are willing to pay for the least attractive of the options.

[17] Rationality does not tell me to choose as a function what will happen *to me* if I do A rather than B (see Chapter 13). It is compatible with some forms of other-regarding morality, but not with the one represented by the categorical imperative.

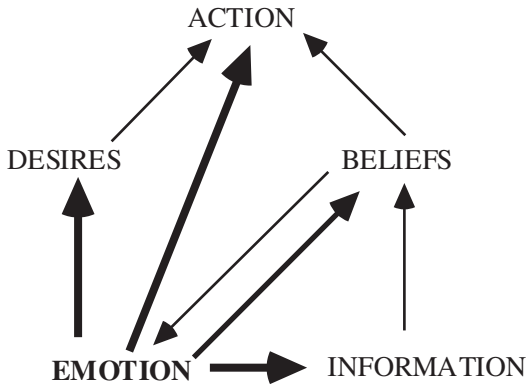



Figure 14.1

tendency they may cause a temporary preference change. If the situation imposes a delay between the time at which the decision to act is made and the time of acting, the action may never be undertaken. An example cited in Chapter 8 was the way the increased expressions of interest after September 11, 2001, in serving in the army did not lead to increased enlistment. If action can be taken immediately, it may sometimes be reversed later when the emotion abates. Thus among the two hundred thousand men who deserted the Union Army during the American Civil War, presumably out of fear, 10 percent returned voluntarily. Second, emotions may induce a temporary change in formal preferences, the rate of time discounting or the degree of risk aversion, thus making previously less preferred options appear to be preferable.

Emotions can affect *beliefs* directly by the mechanisms of wishful thinking and counterwishful thinking (we shall see several examples in Chapter 22 of fear-induced panics). Also, as I noted in Chapter 9, pridefulness or amour-propre may cause one to resist the acknowledgment that one has made a mistake. This may explain, in some cases at least, vulnerability to the sunk-cost fallacy.[18]

Finally, by virtue of the urgency of emotions (Chapter 8) they may interfere with the optimal acquisition of *information* and hence also affect beliefs indirectly. Anger (puzzle 16) and love (18) make us do things we would not have done had we had a more rational policy of information gathering. An observation by Seneca may help us pull some of these strands together: "Anger

[18] In this context, it is interesting that animals do not seem to commit the sunk-cost fallacy: perhaps it is because they do not have any self-image to care about.

is altogether unbalanced; it now rushes farther than it should, now halts sooner than it ought." Often, the urgency induces a neglect of the temporally remote effects of the options, by virtue of the fact that *the determination of long-term consequences is itself time-consuming*. Hence the truncation of the time horizon need not be due to higher discounting of known consequences, but to the fact that some consequences do not even appear on the mental screen of the agent.

**Social norms**. The emotions of contempt and shame play an important role in sustaining social norms (Chapter 21). The rush to vengeance may be due to the urgency of anger, but it could also be due to a social norm that brands as a coward anyone who delays avenging himself. The refusal of duelers to choose the weapons with which they are most proficient is also sustained by the fear of being thought excessively concerned with mere survival rather than with honor. Finally, as Amos Tversky suggested to me, the lawn-mowing puzzle might be explained by the operation of social norms rather than by loss aversion. A resident would not think of mowing his neighbor's lawn because there is a social norm in suburban communities against an adult doing such tasks for money. It simply *is not done*. Voting, too, may reflect the operation of social norms if the act of voting is visible to others and they disapprove of non-voters. Voting in large anonymous elections is more plausibly seen as the result of *moral* norms, which may themselves have irrational aspects (see earlier discussion).

### Bibliographical note

The advice to the ophthalmologist is cited from N. Schrage and F. Kuhn, "Chemical injuries," in F. Kuhn (ed.), *Ocular Traumatology* (New York: Springer, 2008). An outstanding and comprehensive survey of the accumulated anomalies that have undermined basic tenets of rational-choice theory is D. Kahneman, *Thinking Fast and Slow* (New York: Farrar Strauss, and Giroux, 2011). For applications of behavioral economics see D. Hough, *Irrationality in Health Care* (Stanford University Press, 2013) and H. Kunreuther, M. Pauly, and S. McMorrow, *Insurance and Behavioral Economics* (Cambridge University Press, 2013). A large variety of applications are discussed in E. Shafir (ed.), *The Behavioral Foundations of Public Policy* (Princeton University Press, 2012). A critical assessment of rational-choice theory from a perspective somewhat different from the present one is found in D. Green and I. Shapiro, *Pathologies of Rational Choice Theory* (New Haven, CT: Yale University Press, 1994), usefully supplemented by J. Friedman (ed.), *The Rational Choice Controversy* (New Haven, CT: Yale University Press, 1996). For the idea of rationality as a norm, see D. Føllesdal, "The status of rationality assumptions in interpretation and in the explanation

of action," in M. Martin and L. McIntyre (eds.), *Readings in the Philosophy of the Social Sciences* (Cambridge, MA: MIT Press, 1994). I argue for the hyperrationality of the "best interest of the child" principle in Chapter 3 of *Solomonic Judgments* (Cambridge University Press, 1989). A useful discussion of some related issues is J. Wiener, "Managing the iatrogenic risks of risk management," *Risk: Health, Safety and Environment* 9 (1998), 39–82. Most of the subsequent puzzles are discussed in articles reprinted in the source books listed in the bibliographical note to Chapter 7. Exceptions are puzzle 1, for which see A. Blais, *To Vote or Not to Vote: The Merits and Limits of Rational Choice Theory* (Pittsburgh: University of Pittsburgh Press, 2000); puzzle 2, for which see D. Kahneman, "Objective happiness," in D. Kahneman, E. Diener, and N. Schwartz (eds.) *Well-Being: The Foundations of Hedonic Psychology* (New York: Russell Sage, 1999); puzzles 16 and 17, for which see Chapter 3 of my *Alchemies of the Mind* (Cambridge University Press, 1999); puzzle 18, for which see the discussion in Chapter 8; and puzzle 19, for which see L. Lopez, "Two conceptions of the juror," in R. Hastie (ed.), *Inside the Juror: The Psychology of Jury Decision Making* (Cambridge University Press, 1993), pp. 255–62. Most of the alternatives are canvassed in the same source books, except for wishful thinking (Chapter 7), emotions (Chapter 8), and social norms (Chapter 21). It should be noted that the certainty effect is closely related to the very first puzzle that was explicitly presented (in 1953) as a challenge to rational-choice theory, the "Allais paradox." The "Chirac puzzle" is discussed, together with many similar examples, in A. Smith, *Election Timing* (Cambridge University Press, 2004). The claim that animals do not commit the sunk-cost fallacy is defended in H. Arkes and P. Ayton, "The sunk cost and Concorde effects: are humans less rational than lower animals?" *Psychological Bulletin* 125 (1999), 591–600. The example of the jam sales on Broadway is taken from S. Iyengar and M. Lepper, "When choice is demotivating: can one desire too much of a good thing?" *Journal of Personality and Social Psychology* 79 (2000), 995–1006. The reference to embezzlement and theft in old English law is from J. F. Stephen, *A History of English Criminal Law* (London: Macmillan, 1883; Buffalo, NY: Hein, 1964), vol. III, p. 153. The findings about reenlistment in the Union Army are from D. Costa and M. Kahn, "Deserters, social norms and migration," *Journal of Law and Economics* 50 (2007), 323–53. Evidence about heightened time discounting induced by emotion is offered in D. Tice, E. Braslavsky, and R. Baumeister, "Emotional distress regulation takes precedence over impulse control," *Journal of Personality and Social Psychology* 80 (2001), 53–67.

# 15 Responding to irrationality

## Second-best rationality

In the last two chapters I have considered the ideal of rational behavior and the frequent lapses from rationality. These lapses, however widespread and frequent, are not inevitable. If we understand our propensity to make mistakes, we can and do take precautions to make ourselves less likely to make them again, or at least limit the damage if we do. As I have said repeatedly, we *want* to be rational. We may think of these precautionary strategies as a form of imperfect or *second-best rationality*. They should be distinguished from simple *learning*, which occurs when the propensity simply fades away as a result of improved insight. It has been reported, for instance, that when people realize that voting is, in one sense, pointless, they are less likely to vote.[1] Cognitive fallacies that are akin to optical illusions can also be overcome by learning. Just as we learn to ignore the appearance of a stick that looks broken in water, some gamblers presumably learn the hard way that the dice have no memory. I am concerned here, however, with propensities that persist over time.

To cope with our tendencies to behave irrationally, we may use either intrapsychic strategies or extrapsychic devices (*precommitment*). I shall first illustrate how these techniques are used to counteract hyperbolic discounting and the inconsistent behavior it generates, and then discuss their use to control emotional and addictive behavior. These various strategies are not necessarily rational, but many of them are.

## Future selves as allies

An agent who is subject to hyperbolic discounting and knows it is *sophisticated*. Unlike the naive agent who finds himself changing his mind over and over again without understanding the underlying mechanism, the sophisticated agent both is aware of her propensity and deplores it.

[1] Students of economics, in particular, seem to behave in this way.

Anticipating future situations in which she will face the choice between an early small reward and a delayed larger reward, she would like to make herself choose the latter despite her propensity to choose the former. In some cases, she may treat her "future selves" as allies in a common effort to overcome temptations. In other cases, she may treat them as adversaries and try to limit the damage they can do to her "present self." This language, to be sure, is metaphorical, but it will be demetaphorized.

Consider first the case in which the choice between an early small and a delayed large reward can be expected to arise over and over again. The agent can then make herself go for the delayed reward by *bunching* (or *bundling*) the choices.

Let me illustrate by an example from a time I was living up in the hills close to the university where I was teaching. Every day I took my bike to get to campus and back. The return trip involved some steep uphill climbing, so that every day I faced the temptation to get off the bike and walk rather than forcing myself to pedal. When I set out from campus I was firmly committed to staying on the bike all the way, but in the middle of the climb a seductive thought would often occur to me, "Why not walk today, and resume biking tomorrow?" Then, fortunately, a further thought occurred, "What is so special about tomorrow? If I yield to temptation today, does not that predict that I will do so again tomorrow, and the day after, and so on?" The last thought enabled me to stay on the bike.

This intrapsychic device involves a *reframing* of the situation. Rather than thinking about future trips home as involving *a series of choices*, I began to see them as *a choice between two series*: always biking up the hill and always walking the bike. By telling myself that my behavior on one occasion was the best predictor of my behavior on the next occasion, I set up an internal domino effect that raised the stakes and made me go for the delayed reward of improved health rather than for the early reward of relief from discomfort. Referring to Figure 6.1, it can be shown, in fact, that if we place many pairs of rewards identical to A and B on the horizontal line and then form two curves, one for the sum of the present-value curves of all the small rewards and one for the sum of the present-value curves of all the large rewards, the latter curve will lie above the former at the time the first choice has to be made, provided that the number of successive choices to be made is large enough.[2] In other

[2] For an illustration, suppose that the present value of 1 unit of utility at time t in the future is $1/(1 + t)$ and that the agent is twice exposed to the choice between a small reward of 3 and a large of 10. The small rewards become available at times 0 and 6, the large ones at times 3 and 9. At time 0, the present value of the first large reward is $10/(1 + 3) = 2.5$, which is less than the present (instantaneous) value of the small reward. If the choice is made on the basis of this comparison only, the small reward will be chosen. The same choice, for identical reasons, will be made at time 6. As the sum of the present values of the two small rewards is $3 + 3/(1 + 6) \cong 3.43$ and the

words, bundling the choices can make the option of always going for the larger reward preferable to always going for the smaller reward. To be sure, choosing the smaller reward today and the larger reward on all future occasions is even better, but by assumption this option is not in the opportunity set of the agent.[3]

Can this assumption be justified? *Is* my behavior today a good predictor of my behavior tomorrow? In cases that involve a genuine causal effect, this may be true. Biking today will keep my muscles strong so that I can also bike tomorrow.[4] In my case, however, I relied on magical thinking rather than on causal efficacy. Just as many people vote or give to charity under the influence of the thought "If not me, who?" what kept me on the bike was the thought "If not now, when?" Or more elaborately: "There is nothing special about today. If I get off the bike, the causes that made me do it will also operate tomorrow and induce the same behavior. If I do not make an effort now, I never will." In the absence of a genuine causal effect, however, the conclusion does not follow. If I *can* stay on the bike today but decide to get off, I can also stay on it tomorrow. Although false, the reasoning is compelling and, I believe, extremely widespread. It shows that we can enlist one form of irrationality (magical thinking) to combat another (hyperbolic discounting).[5]

To work well, such strategies may have to be framed as binary choices: always doing it or never doing it. For many people, abstention is easier than moderation. Boswell noted that "Johnson, though he could be rigidly *abstemious*, was not a *temperate* man either in eating or drinking. He could refrain, but he could not use moderately."[6] The same problem arises if instead of limiting consumption on each occasion one tries to limit the number of occasions on which one may indulge. The stratagem of laying down in advance what will

sum of the present values of the two large rewards is $10/(1 + 3) + 10/(1 + 9) = 3.5$, bunching will make the agent prefer the two large rewards.

[3] If the agent suffers from "decision myopia" (Chapter 14), the bunching may not work. Suppose, namely, that the agent bundles his choices well ahead of the time when the first small reward becomes available. At the time of bunching, the present value of the stream of large rewards is greater than that of the stream of small rewards, and the agent firmly intends to wait for the first large reward. As he moves closer in time to the moment when the first reward becomes available, this intention may or may not survive. This preference reversal is not due to hyperbolic discounting per se, but to decision myopia.

[4] For an analogy, suppose that by voting I could influence many others, who would otherwise have abstained, to cast a vote as well. Note that on this assumption, no magical thinking is involved, only a causal multiplier effect.

[5] In voting, the effect of magical thinking is to help us overcome the propensity to socially harmful rational behavior rather than to counteract irrationality.

[6] According to Montaigne, the problem is quite general: "It is perhaps easier to do without women altogether than duly and scrupulously to restrict yourself to the company of your wife: a man has more means of living an unworried life in poverty than in duly controlled abundance; behavior governed by reason is more thorny than abstinence." He had a generous conception of reason: "I now defend myself against temperance as I used to do against voluptuousness . . . Wisdom has its excesses and has no less need of moderation than folly."

count as a legitimate occasion is easily eroded. When people resolve not to drink alcohol before dinner, they may find themselves scheduling dinner earlier. The rule of drinking wine only at restaurants, never at home, may cause one to dine out more frequently.[7] Kant's rule of smoking only one pipe after breakfast (Chapter 4) was not unambiguous enough to give him full protection, since as time passed he bought himself bigger and bigger pipes. Similarly, when feasible, the rule "Never do it" may be the only one that can be stably upheld. Since this policy is not feasible with regard to eating, obesity may be more recalcitrant than are addictions to private rules.

The binary choice framing can, however, induce absurdly rigid behavior. Suppose I have told myself never to suffer a single exception to the rule of brushing my teeth every night. On a given occasion I find myself without a toothbrush and decide to walk five miles in a blizzard to buy one. To sustain the decision, I tell myself that if I break the rule on that occasion, I will be on a slippery slope leading to rule violations for ever more trivial reasons, and soon there will be no rule at all and my teeth will fall out. Some people construct very elaborate systems of this kind, in which failure to follow one rule predicts failure with regard to other rules as well, thus raising the stakes even more.[8] Because private rules may have these stultifying effects, they sometimes provide a remedy worse than the disease. In Freudian language (Chapter 4), the rigid impulse control exercised by the superego could do more damage than the impulses from the id.

### Future selves as adversaries

Consider now the case in which the agent confronts the choice between rewards (or punishments) at one of several future dates. (Unlike the previous case, the choice is only supposed to arise once.) The agent may then adopt the intrapsychic device of responding strategically to the known propensity of "future selves" to discount the future hyperbolically. Suppose I am a "hyperbolic procrastinator" who always put things off until tomorrow, and then, when tomorrow arrives, puts them off again until the day after tomorrow. Once I understand that I am subject to this propensity, my optimal behavior changes. Suppose that I can carry out a given unpleasant task in any of three periods, and that the cost of doing so goes up with time. If I am naive, I may tell myself that I will perform the task tomorrow. If I am sophisticated, I know

[7] Chief Justice John Marshall used a more transparent device to get around the rule that the court would indulge in drinking wine only when it was raining. Looking out of the window on a sunny day, he would say that "our jurisdiction extends over so large a territory that the doctrine of chances makes it certain that it must be raining somewhere."

[8] They may believe, erroneously (Chapter 12), that cross-situational consistency will induce cross-situational triggering of breakdowns.

that tomorrow I will delay until the last period. The understanding that the cost will in fact be very high unless I perform the task right away may induce me to do exactly that.[9]

In this case, being sophisticated helps. In other cases, being naive may be better. Suppose you can have a reward in any one of three successive periods and that the rewards increase with time. An example might be a person who has been offered a bottle of wine that improves with time up to the third year, and then deteriorates. A naive person may form the intention to wait until the third period, and then change her mind and drink the wine in the second period. A sophisticated person will know that he is never going to wait until the third period, so that he effectively only faces the choice between the first-period reward or the second-period reward. In that choice, the early reward may win out.[10] Some alcoholics report being subject to a similar kind of reasoning: "I know I am going to yield to temptation, so I might as well do it right away." Also, naive smokers who quit for the first time may hold out longer than sophisticated smokers who have tried several times and know the odds against succeeding. Although backsliding in addiction need not be due to hyperbolic discounting, the general point is the same: if you can predict that you will deviate from your best plan, you may end up deviating even more from it or earlier than if you are unaware that you will fail.

### New Year's resolutions

The theory of focal points (Chapter 18) predicts that people will try to quit a bad habit or an addiction on some salient day, such as January 1, to resist the temptation to say to oneself, "Why not wait until tomorrow?" In one study, 213 participants made an average of 1.8 New Year's resolutions. Smoking

[9] For an illustration, suppose that the present value of 1 unit of utility at time t in the future is $1/(1 + t)$ and that I will suffer increasingly by delaying my visit to the dentist: if I go today I suffer a pain of –2.75, tomorrow it will be –5, and the day after –9. From today's perspective, the present values are, respectively, –2.75, $-5/(1 + 1) = -2.5$, and $-9/(1 + 2) = -3$. Hence it might seem that the optimal choice from today's perspective is to postpone the visit until tomorrow. However, being sophisticated I know (today) that tomorrow the present value of going tomorrow will be –5 and that of going the day after $-9/(1 + 1) = -4.5$, inducing a preference to wait until the day after. Since today, however, I prefer going today to going the day after tomorrow, I go today.

[10] For an illustration, suppose that the present value of 1 unit of utility at time t in the future is $1/(1 + t)$ and that I can benefit more and more by delaying my consumption of a bottle of wine: if I drink it this year I derive a pleasure of 2.75, next year it will be 5, and the year after 9. From today's perspective, the present values are, respectively, 2.75, $5/(1 + 1) = 2.5$ and $9/(1 + 2) = 3$. Hence it might seem that the optimal choice from the first year's perspective is to drink the wine the year after next. However, being sophisticated I know (this year) that next year the present value of drinking it next year will be 5 and that of drinking it the year after $9/(1 + 1) = 4.5$, inducing a preference to drink it the next year. Since in the first year, however, I prefer drinking it the first year to drinking it next year, I drink it today.

cessation (30 percent) and weight loss (38 percent) together accounted for two-thirds of the resolutions. Relationship improvement (5 percent), reduction in alcohol consumption (2 percent), and an increase in monetary savings (2 percent) were also popular resolutions. The "other" category, representing 23 percent of the primary pledges, contained a multivarious range of idiosyncratic responses, such as temper control, setting aside time for oneself, making decisions oneself, and learning to say no. Although the study reports success rates and suggests some explanatory variables, it does not say anything about the efficacy of the New Year's resolutions compared to the decision to stop on any other day.

There is some indirect evidence, however, for their efficacy: cigarette companies seem to believe in it. If most people try to quit on January 1 and, as the evidence suggests, withdrawal symptoms peak after about one month, quitters should be particularly vulnerable to cue-triggering in January and February. One might expect, therefore, that there would be a peak in cigarette advertising in those months, an expectation confirmed by an analysis of the back covers of 3,024 magazines. After considering and refuting three alternative explanations, the authors conclude that the timing of the advertisements was deliberately chosen to counter New Year's resolutions. The conclusion is certainly consistent with what we know about the culture of hypocrisy and manipulation in American cigarette companies, second only to that of the National Rifle Association.

The findings suggest a three-level model. At the first level, there is the desire of an individual to smoke or engage in other excessive (legal) behaviors. At the second level, there is the desire of an individual who engages in these behaviors, to quit them. At the third level, there is the desire of the companies who profit from the behaviors to prevent the consumers from quitting. Objectively, there is an alliance between the first and the third level. The desire of the user in a weak moment and the desire of the companies are identical. A person who wants to quit is squeezed in the middle.

### Extrapsychic devices

In practice, intrapsychic devices may be less important than the precommitment devices to which I now turn. These involve affecting the external world, in ways that cannot be instantly and costlessly undone, for the purpose of making it less likely that one will choose the earlier, smaller reward in the future. Six strategies stand out: *eliminating* the choice of the early reward from the feasible set, *imposing a penalty* on the choice of the early reward, *adding a premium* for the choice of the delayed reward, *imposing a delay* between the choice and the actual delivery of the reward, *avoiding cues* that might trigger preference reversal, and *avoiding information* that might

cause you to choose the immediate reward. Saving behavior can illustrate the first four options. If I begin saving for Christmas but find myself taking money out of my savings account instead of keeping it there, I may join a Christmas savings club that will not allow early withdrawals (Chapter 14). Alternatively, I may put my savings into a high-interest account that carries a penalty for early withdrawal, thus combining premium and penalty. If I want to save for my old age, I may set up a delay between the time I might make a decision to dissave and the moment the funds become available, by investing in illiquid assets rather than in stocks or bonds.

The fifth option is illustrated by the person whose craving for dessert is triggered by visual cues. The trick is to go to a restaurant where they do not wheel around the dessert trolley, so that one has to order from the menu. We may contrast this with a person who has a dessert problem because of hyperbolic discounting. For him, the best option is to go a restaurant where he has to order dessert at the beginning of the meal. Both strategies involve protecting oneself against the effects of the *proximity of temptation*, be it spatial or temporal.

To understand the sixth option, consider a person who is considering having unprotected sex. On the basis of her approximate knowledge about the risks of AIDS transmission, she decides to abstain. She knows that she could easily obtain more accurate information, but abstains from doing that too, on the basis of the following reasoning. If it turns out that the risk is smaller than she thought, she might then decide to have unprotected sex, even though doing so would be suboptimal both from her present point of view and from her future point of view with less accurate information. (Here, I assume that she discounts the future hyperbolically.) It is not clear how often people actually go through this chain of reasoning in a conscious manner. Strategic ignorance, such as avoiding going on the scales, is probably more common when people want to *persist* in harmful behavior.

People who sign up for weekly physical exercises often drop out after a week or two. To prevent this, they may (in theory at least) sign a contract with the fitness center to pay twice the normal fee up front and receive a fraction back each time they show up. People who sign up for weight loss programs may have to pay a deposit that they get back only if they lose a stipulated amount of weight, sometimes with the rider that if they fail, the deposit will be donated to the person's most disliked political cause. When I set out to give the lectures that resulted in the first edition of the present book, I precommitted myself by telling my students that I would give them a draft chapter at the end of each week. If I had failed to live up to that promise, I would have suffered the cost of their mild ridicule. If I am afraid I might cancel my appointment with the dentist when the time approaches, I can authorize him to bill me twice the regular amount if I fail to show up. In the case of wine that will improve

with time, you may ask the seller to store it for you to protect yourself against premature gratification. If you are afraid that you might read the last novel by your favorite crime writer too quickly, skimming paragraphs to get to the dénouement, you might buy a book-on-tape version (and a player with no fast-forward function) that leaves you no choice but to listen to every word.

The examples in the last two paragraphs involve precommitment against two kinds of temptation. On the one hand there is *procrastination*, including failures to save, to seek painful treatments, to do physical exercise, or to write up a manuscript. On the other hand, there is *premature gratification*, such as drinking wine too early or skipping pages in a book. These temptations stem directly from hyperbolic discounting. Nothing but the sheer passage of time is involved. In a further category of cases, *excessive behavior*, hyperbolic discounting may interact with other, visceral motivations. These include overeating, compulsive gambling, and addictive behavior. It may be hard to know, in such cases, whether preference reversal is due to the discounting structure or to other factors. A decision to fast that is made on a full stomach may dissolve as the person again begins to feel hungry. A decision to stop smoking may be eroded by the sight of another person lighting up a cigarette. This is the phenomenon of *cue dependence* – cravings that are triggered by visual cues associated with the consumption of the addictive substance. A decision to stop gambling that is induced by the guilt feelings of the gambler toward her family may unravel once the emotion fades in strength (Chapter 8). It may also be hard to tell whether we are dealing with procrastination or with visceral factors. A decision to take medication regularly may be undermined by the decline of the strong emotions that caused the patient to see a doctor in the first place.

Once the agent understands that he is subject to the latter mechanisms, he may precommit himself to forestall their operation. To prevent his resolve to diet from being undermined by hunger, he may take a pill that attenuates the craving for food. More drastically, he may have his jaws wired so that he can only take in liquid sustenance. If he knows that his desire for dessert is cue dependent, he will not go to restaurants where they present a dessert trolley. Former heroin addicts will stay away from the places where they used to consume the drug. Ex-gamblers learn not to go to a casino "just to watch others play." If the agent can predict that her anger will fade so that she will not want to punish the offender, she might carry out the punishment immediately. As noted earlier, this behavior was observed in Belgium after 1945.

In fighting addiction, the strategy of imposing costs on oneself is very common. When General de Gaulle wanted to quit smoking, he told all his friends about it, to increase the costs of backsliding. In his case, the reputation loss would have been very high. In a cocaine addiction center in Denver, doctors are offered the opportunity to write a self-incriminating letter to the State Board of Medical Examiners confessing to drug use and asking that their

license be revoked. The letter is automatically mailed if the patient tests positive for cocaine use. Some former alcoholics try to stay dry by taking disulfiram, a drug that has the effect of making the user violently ill if he takes a drink.[11]

Self-imposed delays can also be effective in resisting cravings. To prevent myself from impulse drinking, I may store my liquor in a safe with a timing device. Alternatively, I may adopt a policy of having no liquor at home so that I have to go to a store to get it. The disulfiram technique in fact combines the imposition of costs with delays, since once you have taken the pill you have to go two days without taking it before you can drink without getting sick. The cocaine addiction center, too, combined costs with delays. It allowed people to break out of the compact by submitting a notarized declaration of withdrawal from the arrangement. There was a two-week delay. Anyone who submitted a request for withdrawal could retrieve the incriminating letter after two weeks. But if during the two weeks' interim, the withdrawal was rescinded, then it would require another two weeks' notice. Although many of the patients invoked the withdrawal procedure, none went two weeks without revoking the revocation.

The concern with precommitment against time inconsistency and excessive behavior is relatively recent. The classical writers on the topic focused on precommitment against *passion*, taken in a wide sense that also includes intoxication and psychotic states.[12] In the *Odyssey*, Homer offered what has become the standard example of precommitment: Ulysses binding himself to the mast so that he would be unable to respond to the song of the Sirens. In *On Anger*, Seneca wrote, "While we are sane, while we are ourselves, let us ask help against an evil that is powerful and oft indulged by us. Those who cannot carry their wine discreetly and fear that they will be rash and insolent in their cups, instruct their friends to remove them from the feast; those who have learned that they are unreasonable when they are sick, give orders that in times of illness they are not to be obeyed." In Mme de Lafayette's novel *La Princesse de Clèves*, the princess flees the court to avoid the temptation of responding to the overtures of the Duc de Nemours; even later, when her husband is dead and she is free to remarry, she stays away. "Knowing how circumstances affect the wisest resolutions, she was unwilling to run the risk of seeing her own altered, or of returning to the place where lived the man she had loved." In Stendhal's novel *Lucien Leuwen*, Mme de Chasteller takes care to

[11] The most common form is by oral intake, which works by causing the consumption of alcohol to make you sick. The (physically inefficacious but psychologically efficacious) implantation under the skin is less common.

[12] The phenomenon, briefly mentioned in the text, of precommitting oneself while being in the grip of passion and fearful that it will abate is much less common.

see Lucien only in the company of a chaperone, to make it prohibitively costly to give in to her love for him.

These strategies are quite common. When people burn their bridges, it may be for strategic reasons (Chapter 18), but probably more often to prevent themselves or others from giving in to fear. I may stay away from the office party because I know from past experience that I am likely to have a drink or several, and that because of its disinhibitory effect alcohol will induce aggressive or amorous behavior that I will later regret. Alternatively, I may decide to take my spouse along, to raise the cost of such behavior. Merely *resolving* not to drink (or not to get emotional if I do) is less likely to be effective, given "the power of the situation" (Chapter 12). Similarly, controlling anger by the intrapsychic device of counting to ten before talking back or lashing out presupposes a detachment that tends to be lacking in the heat of the moment. A general advice of self-help is in fact to "break the chain early" rather than to rely on self-control in the face of temptation or provocation. As Mark Twain said, "It is easier to stay out than get out." For an extreme case, I refer to a *New York Times* headline (April 5, 1996): "Texas Agrees to Surgery for a Molester: Soon to Leave Prison, Man Wants Castration to Curb His Sex Urge."

Delay strategies might seem to hold out the best promise for dealing with emotion-based irrationality. Since emotions tend to have a short half-life, any obstacle to the immediate execution of an action tendency could be an effective remedy. As I note later, public authorities do indeed count on this feature of emotion when they require people to wait before making certain important decisions. It is rare, however, to observe people imposing delays on *themselves* for the purpose of counteracting passion. The requisite technologies may simply be lacking. One example, however, is the "covenant marriage" offered by three American states (Arkansas, Arizona, and Louisiana), an optional form of marriage that is harder to enter and harder to leave than the regular marriage. Typically, a couple who have entered a covenant marriage can be granted a divorce only after two years of separation, as compared to a normal waiting time of six months. The small minority (less than 1 percent of marrying couples) who use this option presumably do so to signal their commitment to each other and to protect themselves against short-lived passions and temptations.

Precommitment often involves the help of other individuals, organizations, or public authorities. These need, however, to be independent from the agent issuing the precommitment instructions, since otherwise he might revoke them. To fight his opium addiction, Samuel Coleridge hired a man to oppose by force his entrance into any druggist's shop. When the man tried to restrain him, Coleridge said, "Oh, nonsense. An emergency, a shocking emergency has arisen – quite unlooked for. No matter what I told you in times long past. That which I *now* tell you, is – that, if you don't remove that arm of yours from the

doorway of this most respectable druggist, I shall have a good ground of action against you for assault and battery." Similarly, Mao Zedong gave orders that any orders he might issue after taking sleeping pills were to be ignored. When, after having taken the pills, he ordered his aide to send an invitation to the American table tennis team to visit China (the beginning of Chinese-American relations) and the aide asked him, "Do your words count?" he answered, "Yes, they do. Do it quickly. Otherwise there won't be time."

Organizations are more reliable tools for self-binding. The cocaine clinic in Denver and the Christmas clubs offer self-binding options that the individuals could not have come up with on their own and that were deliberately designed to help them to overcome their problems[13] *and* to prevent them from rescinding their instructions. In Norway, the Law of Psychic Health Protection allows individuals to commit themselves v*oluntarily but irreversibly* to a three-week treatment in a psychiatric institution. It seems, however, that the system does not work, because doctors have the right but not the duty to retain individuals once they are in the clinic. To make it effective, patients would have to be allowed to sue their hospital if, at their request, it released them prematurely.

In 1996, the state of Missouri began a self-exclusion program for compulsive gamblers. Anyone who signs up for a self-exclusion list is banned for life from entering any of Missouri's casino riverboats. If she tries to ignore the ban and gamble on one of Missouri's riverboats anyway, she is to be removed from the boat, and "the licensee shall cooperate with the commission agent in reporting the incident to the proper prosecuting authority and request charges be filed ... for criminal trespassing, a class B misdemeanor." Self-excluded gamblers are to be denied any winnings if they somehow manage to go aboard a riverboat, gamble, and win.

The state can also take a more active role, by imposing delays on abortion, gun purchase, or divorce (and marriage!) and by allowing consumers a three-day or week-long cooling-off period during which they can cancel purchases made in a moment of enthusiasm. To illustrate, consider the 238,292 purchasers of handguns in California in 1991. In the year following the purchase, 21.9 percent of the deaths among the purchasers were suicides by firearm, as against 0.9 percent in the general population. The temporal profile of the suicides among the purchasers illustrates strikingly the short half-life of emotions (Chapter 8): after the obligatory waiting time of fifteen days, the number of suicides in the first week was double that of the fourth, the number in the first month five times that of the twelfth, and the number in the first year six times that of the sixth. Had there been no waiting time, there would presumably have been more suicides; with a longer wait, fewer.

[13] Safes with timing devices, by contrast, were not made to help people fight their drinking problems.

Softer methods are also used. At Gardermoen airport in Norway, tobacco products are no longer on open display in the tax-free shop, but shuttled away in a separate room, presumably to prevent cue-triggered cravings in smokers who are trying to quit.[14] In Norwegian bars, you cannot order a double whisky (8 cl), but nothing prevents you from ordering two singles in succession.[15] When Mayor Bloomberg tried to impose a similar limitation on the size of soda bottles, he failed.

Sometimes, political constitutions are understood as precommitment devices, or a form of *collective self-paternalism*. John Potter Stockton, writing in 1871, said that "constitutions are chains with which men bind themselves in their sane moments that they may not die by a suicidal hand in the day of their frenzy." Another common metaphor is that constitutions are ties imposed by Peter when sober on Peter when drunk. Bicameralism is often cited as an example of political precommitment: by having all legislation pass through two houses, one creates time for impulsive passions to cool down and reason (or interest!) to regain the upper hand. In Chapter 25, I make some skeptical comments on this argument. Imposing delays on constitutional amendments has been justified by the same argument. If precommitment is understood as *self*-binding, however, the extension from the individual to the collective case, and from the intragenerational to the intergenerational case, is quite dubious. Rather than a community's binding itself, we find majorities binding minorities and the present generation binding the future. Moreover, since constitutions are typically written in turbulent times, framers or founders are often themselves in the grip of passion. Being "drunk," they may not see the need to take precautions against drunkenness. Thus on September 7, 1789, when the French Assemblée Constituante was debating whether to write unicameralism or bicameralism into the constitution, the deputy Adrien Duquesnoy wrote the following entry into his journal: "If one can be allowed to make a probability assessment, it seems clear that the majority of the assembly will never vote for the two chambers. This outcome may have great disadvantages, but the situation is such, and the minds are so exalted, that no other is possible; perhaps it will be possible to make a change in a few years. One will come to understand that a unique assembly, in a nation as extremely impetuous as ours, can produce the most terrible effects."

[14] The importance of cue-triggering was observed in Sweden, where sales of alcohol went up by 10 percent after a change from over-the-counter to self-service sales.

[15] When he was a political commentator, my father argued, unsuccessfully, that the government should ban advertising that offers products at the price of $499.95 and similar prices slightly below a round number, because they exploit consumer irrationality.

### Bibliographical note

In this chapter I draw on my book *Agir contre soi* (Paris: Odile Jacob, 2007). The intrapsychic device of bundling or bunching the options has been extensively discussed by G. Ainslie, notably in *Picoeconomics* (Cambridge University Press, 1992). The information-avoidance strategy was proposed by J. Carrillo and I. Mariotti, "Strategic ignorance as a self-disciplining device," *Review of Economic Studies* 67 (2000), 529–44, and confirmed experimentally by T. Brown, R. Croson, and T. Eckel, "Intra- and inter-personal strategic ignorance: a test of Carrillo and Mariotti" (accessible at http://stiet.cms.si.umich.edu/node/809). The discussion of decision myopia draws on O.-J. Skog, "Hyperbolic discounting, willpower, and addiction," in J. Elster (ed.), *Addiction*: *Entries and Exits* (New York: Russell Sage Foundation, 1999). Strategic responses by sophisticated individuals who are aware of their propensity to discount the future hyperbolically are discussed by T. O'Donoghue and M. Rabin, "Doing it now or later," *American Economic Review* 89 (1999), 103–24. The study of New Year's resolutions is J. Norcross, A. Ratzin, and D. Payne, "Ringing in the new year: the change processes and reported outcomes of resolutions," *Addictive Behaviors* 14 (1989), 205–12, and the study of the countermeasures by the cigarette companies is M. Basil, D. Basil, and C. Schooler, "Cigarette advertising to counter New Year's resolutions," *Journal of Health Communication* 5 (2000), 161–74. The idea of precommitment or self-binding to cope with one's irrational propensities is discussed in T. Schelling, "Egonomics, or the art of self-management," *American Economic Review: Papers and Proceedings* 68 (1978), 290–4, and in several of his later publications. I have discussed it in *Ulysses and the Sirens*, rev. edn (Cambridge University Press, 1984); in *Ulysses Unbound* (Cambridge University Press, 2000); and in "Don't burn your bridge before you come to it: ambiguities and complexities of precommitment," *Texas Law Review* 81 (2003), 1751–88. A book-length treatment of the failure to take prescribed medications is G. Reach, *Pourquoi se soigne-t-on?* (Paris: Éditions de Bord de l'Eau, 2005). The story about Coleridge is found in Thomas de Quincy, *Confessions of an Opium Eater* (London: Penguin, 1968), p. 145. The story about Mao is found in J. Chang and J. Halliday, *Mao: The Unknown Story* (New York: Knopf, 2005), 580–1. The increase in the sales of alcohol in Sweden is documented in O.-J. Skog, "An experimental study of a change from over-the-counter to self-service sales of alcoholic beverages in monopoly outlets," *Journal of Studies on Alcohol* 61 (2000), 95–100. The data on handguns and suicides in California are from G. Wintermute *et al.*, "Mortality among recent purchasers of handguns," *New England Journal of Medicine* 341 (1999), 1583–9. Whereas in *Ulysses and the Sirens* I was enthusiastic about the idea of constitutions as precommitment devices, I recanted in *Ulysses Unbound*.

# 16 Implications for textual interpretation

In a common view, the scientific enterprise has three distinct parts or branches: the humanities, the social sciences, and the natural sciences.[1] For some purposes, this is a useful way of carving up the field of science, but for other purposes a rigid distinction may prevent cross-fertilization. In this chapter I argue that the humanities and the social sciences have more in common than is usually assumed. (In Chapter 11, I argued that the relevance of natural sciences for the study of society, while not nil, is more limited.) In particular, I shall try to show that *interpretation* of works of art and *explanation* are closely related enterprises. In one sense, this seems to me to be self-evident. The production of a work of art is an *action*, or a series of actions. Like any other action, it is in principle susceptible to explanation in terms of the antecedent mental states of the agent, be they conscious or unconscious. This view of interpretation has the advantage that there is a *fact of the matter* by virtue of which the interpretation is right if it is right, and wrong if it is wrong. By contrast, interpretations that focus only on "the work itself" have no such external criteria of adequacy. Our proneness to autosuggestion (see Introduction to Part II) and our propensity to search for objective teleology and analogy (Chapter 9) may combine to yield arbitrary interpretations.

A *successful* work of art is one that can be given a rational-choice explanation, in the sense that the author's choice of means is adequate for his ends. Once again, this seems to be a self-evident proposition. In practice, however, we may not be able to verify this adequacy, except negatively. We may be able to identify a jarring note, but not to explain the choice between several aesthetically plausible options. Even identifying jarring notes – irrational choices – may be impossible if the jarring character itself is sought as an

[1] Based on casual observations of a few elite academic institutions, I conjecture that in their prestige hierarchy natural scientists and mathematicians come first, next scholars in the humanities, with social scientists at the bottom. For members of the first group, who set the tone, members of the second display admirable qualities of erudition or eloquence, whereas members of the third are merely engaged in a poor imitation of what they do themselves.

aesthetic effect.[2] For this reason, I shall limit myself to works of art where the criteria are relatively unambiguous: classical (pre-1850) novels and plays defined by the tacit convention that the events and characters that are described *could have been real*.[3]

Consider first rationality as a motive of the *characters* in fiction or plays. A classical problem in literary criticism is why Hamlet delays taking revenge for his father's death. Many explanations have been offered. Some of them appeal to irrationality, in terms of weakness of will or clinical depression. There is, however, also a simple rational-choice account. Although Hamlet initially believes what his father's ghost told him about Claudius, he later decides to *gather more information* by staging a play to "catch the conscience of the king." Once the reactions of the king have confirmed his belief, however, he *lacks an opportunity* to realize his desire, which is to make Claudius burn in hell forever. Although he has an opportunity to kill Claudius while he is praying, doing so would, according to contemporary theology, ensure Claudius salvation rather than damnation. Later, he kills Polonius behind a curtain, *wrongly but not irrationally* believing him to be the king. Given the information he had, his belief that it was the king hiding behind the curtain was rational. Moreover, he had no reason to gather *more* information, since he could reasonably assume that someone hiding behind the curtain in the queen's presence would be the king.

I do not claim that this is the right interpretation (in fact I have not yet said what it means for an interpretation to be "right"). My point is simply that the three episodes I have mentioned are prima facie consistent with the idea that Hamlet is rationally pursuing the goal of avenging his father's murder. Another question is whether the idea is consistent with Hamlet's repeated self-accusations for lacking the resolve to take revenge. Many commentators interpret these famous monologues as a sign of weakness of will and view the two first episodes as based on self-deceptive excuses for inaction. (The third episode is harder to square with this view.) Now, although weakness of will and self-deception violate the canons of rationality, they are perfectly *intelligible* (Chapter 3). When dealing with the internal development of the work of art, intelligibility rather than rationality is the most useful idea for the task of interpretation.

In contrast to the internal point of view, we may take the external point of view of the author. To the question "Why does Hamlet delay his revenge until

[2] As stated on the label of a denim jacket I bought in San Francisco some 40 years ago, "Any defect or fault in this garment is intentional and part of the design."

[3] An early example of a violation of this convention occurs toward the end of Ibsen's *Peer Gynt* (1867), when Peer is afraid of drowning and the "strange passenger" tells him that "one does not die in the middle of the fifth act."

the fifth act?" we might answer, "The death of the king must take place at the end of the play."[4] This is a matter of dramaturgical construction, not of psychology. By itself, this answer would not be satisfactory. If Shakespeare had dragged out the revenge by a series of arbitrary events or ad hoc coincidences, simply for the purpose of having it occur at the end of the play, we would have deemed it an authorial failure. More pointedly, it would have been a case of *authorial irrationality*.

Authorial rationality is like the rationality imputed to God. Like God, the author is setting in motion a process in which each event can be *explained twice over*, first causally and then teleologically. I take this idea from Leibniz, who wrote that there are

> two kingdoms, one of efficient causes, the other of final, each of which separately suffices in detail to give a reason for the whole, as if the other did not exist. But neither is adequate without the other when we consider their origin, for they emanate from one source in which the power that makes efficient causes, and the wisdom which rules final causes, are found united.

God's aim is to create the best of all possible worlds. Specified to include the temporal dimension, the idea can be understood as the *best of all possible sequences*. Although the transition from one state of the universe to the next occurs by ordinary physical causality, the initial state and the laws of causality have been chosen so as to maximize the overall perfection of the sequence.

If we limit ourselves to the classical drama or the classical novel, the author's task is to develop the plot through what the characters say and do, often in response to one another. The aim is to do so in a way that maximizes aesthetic value. Thus each action or statement by a character can be explained twice over, both as a reaction to previous actions and statements (or external events) and as a generator of surprise, tension, and ultimately tension resolution in the reader. Here is an example offered by a literary theorist: "Suppose we want to know 'why' in the early part of Dickens's *Great Expectations* . . . the six- or seven-year old Pip aids the runaway convict. Two different kinds of answer are possible: (1) according to the logic of verisimilitude (made prominent, in fact, by the text): the child was frightened into submission; (2) according to the structural needs of the plot: this act is necessary for Magwitch to be grateful to Pip so as to wish to repay him; without it the plot would not be the kind of plot it is."[5]

[4] Unlike the words Ibsen puts in the mouth of "the strange passenger" (see previous note), Shakespeare could not have had Hamlet say, "I cannot kill the king until Act V."

[5] The passage is strikingly similar to Leibniz's assertion. According to a novelist who was also an accomplished theologian, the meaning of the biblical phrase that God created Man in his own image is that God created Man with the desire and ability to create.

The fact that authors (and other artists) often make many drafts before they are satisfied, or before they lay down their pens or brushes, is irrefutable evidence that they are engaged in a process of *choice* and that they possess explicit or implicit criteria for *betterness*. Proust wrote sixteen drafts of the opening chapter of *À la recherche du temps perdu*. Picasso filled sixteen sketchbooks with drawings in preparation for "Les Demoiselles d'Avignon." The fact that these drafts typically involve *small* variations suggests that they are aiming at a *local maximum* of whatever form of betterness they are striving for. However, the difference between an author and someone who is merely climbing along a gradient is that the former's *creativity* goes beyond mere choice (Chapter 13). The reason why the creation of a work of literature cannot be reduced to rational choice is that the number of meaningful word sequences is too large for one person to scan them all and select "the best." Although a "rational creator" may try to make the problem more tractable by deliberately excluding some sequences (as noted in Chapter 10, this is one of the functions of meter and rhyme in verse), too many options will usually remain for choice to be a feasible selection mechanism. Instead, the author will have to rely on his or her unconscious associative machinery.

Rational creation is therefore largely about getting the second decimal right or, to shift the metaphor, about climbing to the top of the nearest hill. The task of finding a hill that towers over the others is not within the scope of rationality. Yet even reduced to the task of fine-tuning, authorial rationality matters. As suggested by the phrase "a minor masterpiece," it may be better to find the top of a low hill than to remain on the slopes of a taller one. Without implying any comparative judgment, *Chronicle of a Death Foretold* and *Look Homeward, Angel* can serve to illustrate the two possibilities.

Let me enumerate and then discuss some demands that rationality imposes on the author. First, the acts and utterances of the characters have to be intelligible. Second, the author has to meet the twin requirements of *fullness* and *parsimony*. Third, the work has to flow *downhill*, in the sense of minimizing the appeal to accidents and coincidences. Fourth, it has to offer a psychologically gratifying pattern of the buildup and resolution of tension.

Intelligibility can be absolute or relative, and if relative, global or local. The question of absolute intelligibility is whether *any* human being could behave in this way. The question of relative global intelligibility is whether the behavior of a fictional person is consistent with his or her overall character as displayed earlier in the work. The question of relative local intelligibility is whether the behavior of a fictional person is consistent with his or her behavior in similar situations earlier in the work. Whereas the requirements of absolute and of relative local intelligibility are crucial constraints on authorial rationality, that of relative global intelligibility is not. If anything, the respect for the latter constraint may be seen as an aesthetic flaw.

In some cases, absolute intelligibility may be violated by excess of rationality. Consider again Euripides' Medea or Racine's Phèdre, both equally lucid about their self-destructive passions. They are portrayed as being subject to weakness of will in the strict sense, knowing that what they are doing is contrary to the all-things-considered judgment they hold *at the very moment of acting*. Although passion causes them to deviate from that judgment, it does not affect it. Racine's Hermione is a more credible character. Because her judgment is clouded by her emotions, she is self-deceptive rather than weak willed. My suggestion – it is nothing more than that – is that the simultaneous presence of extreme emotion and full cognitive lucidity goes against what we know about human nature.

Whereas too much rationality can be unintelligible, irrationality can be perfectly intelligible. What can be more intelligible than the reaction of M. de Rênal in Stendhal's *Le rouge et le noir* when, in the face of strong signs that his wife is having an affair with Julien Sorel, he chooses to believe in her fidelity? The wish is the father of the thought. More paradoxical are cases in which the desire that one's wife be faithful causes the belief that she is *not*, against the evidence. In *Othello*, "Trifles light as air are to the jealous confirmation strong as proofs from holy writ." The first is a case of short-circuiting, the second one of wire crossing (Chapter 3).

Relative intelligibility, which is violated by a person in a play or a novel who acts "out of character," raises different problems. First, we must take account of arguments by psychologists that character traits tend to be *local* rather than global (Chapter 12). Whereas many authors (Hamsun mentions Zola) subscribe to the folk psychology that assumes cross-situational consistency, good authors (he mentions Dostoyevsky) do not. The latter may disappoint readers who expect characters to behave "in character," but these are not the intended audience of the work. As we shall see shortly, even good authors may be constrained by the flawed psychology of their readers, but the belief in global traits is not one they should respect. Readers have a right, however, to expect local consistency. If the author paints himself into a corner, so that the only way to develop the plot as planned is to allow for a character to act in a locally inconsistent manner, he is violating his implicit contract with the readers. A plot should develop as water seeks its natural downhill course, not by the author's forcing it to run uphill.

Let me illustrate this idea by some of Stendhal's marginal comments in the manuscript of his unfinished and posthumously published novel *Lucien Leuwen*. Stendhal has the eponymous hero fall in love with a young widow, Mme de Chasteller. His feelings are reciprocated, but he does not dare to reach out to her. The very delicacy of mind that makes him superior to "the most accomplished Don Juan" and hence capable of inspiring love also makes him inferior to any "less well-bred young Parisian" who would instantly know how to

handle the situation. To move the plot forward, Stendhal needs to bring them together but does not quite know how to do it. He writes in the margin: "Upon which the chronicler says: one cannot expect a virtuous woman to give herself absolutely; she has to be taken. The best hunting dog can do no more than bring the game within gunshot. If the hunter doesn't shoot, the dog is helpless. The novelist is like the dog of his hero." The comment strikingly illustrates the need for the behavior of characters in a novel to be "in character."

Stendhal does eventually manage to engineer a situation in which the love of Lucien and Mme de Chasteller for each other can be shown and understood, and yet not be declared. But his difficulties do not end there. Stendhal's plan for the novel followed the dialectical Hollywood recipe: boy meets girl, boy and girl break up, boy and girl reunite. As we just saw, he had problems getting the thesis established. To produce the antithesis, Stendhal uses the ridiculous and manifestly teleological device of making Lucien believe that Mme de Chasteller, whom he has seen daily at close quarters, has suddenly given birth to a child. But what really stumped him was the synthesis. Although we do not know why he never got around to writing the third part in which the lovers would be reunited, one conjecture is that their union would not be plausible. In the second part of the novel, after the breakup, Lucien turns into a bit of a cynical rake, fundamentally honest by the lax standards of the July Monarchy but certainly very different from the awkwardly delicate person with whom Mme de Chasteller had fallen in love. Stendhal may have decided that having her love the transformed Lucien would violate relative intelligibility.

Aristotle wrote that "the story ... must represent one action, a complete whole, with its several incidents so closely connected that the transposition or withdrawal of any one of them will disjoint and dislocate the whole. For that which makes no perceptible difference by its presence or absence is no real part of the whole." We may read this passage as expressing the two aesthetic ideals of *fullness* and *parsimony*. The reader is entitled to think that the author has presented her with all the information she needs to understand the development of the plot.[6] Conversely, she is entitled to expect that if the author tells her that it was raining when a character left his house, it is because the premise of rain will be needed later on, and entitled to believe that a speech attributed to

[6] To be sure, potentially relevant details may deliberately be left out to leave some room for the imagination of the reader. Rational creation is compatible with (and may even demand) some blanks to be filled out by the reader. (That may be one reason why the movie version of a novel is often less satisfactory than the book, and why seeing the movie first may detract from the pleasure of reading the novel.) If, however, the artist overestimates the imagination of her audience, her effort will be deemed a failure. Suppose a novelist tries to suggest the temperamental incompatibility of a hero and heroine by making the street numbers of the houses in which they live mutually prime, that is, having no common divisors. Barring special circumstances, she cannot reasonably count on the reader's being able to pick up that fact.

a character is intended to tell us something about the person or to serve as a premise for the action of other characters.[7]

Earlier, I referred to the "downhill" character of a good plot, using acting "in character" as an example. More generally, good plots should not turn on unlikely events, accidents, and coincidences. In *Middlemarch*, the encounter between Raffles and Mr. Bulstrode – a crucial element in the development of the story – is so contrived that it detracts from the otherwise seamless progression of the novel. Accidents may, to be sure, have their place in a novel. The accidental death of a parent may trigger or shape the unfolding of a plot, as may the death of both parents in the same accident. But if the plot requires their deaths in *two* separate accidents, credulity is strained. The convenient death of a spouse that allows the hero or heroine to marry his or her real love is also a sign of blamable authorial laziness. The introduction of twins in a detective novel when the author has painted himself into a corner is an even more blatant failure.

The psychology of readers is not, however, finely attuned to probability theory. Suppose the author has the choice between getting from A to B in a plot in two steps or in six steps. For specificity, suppose that the two steps require events that will occur with likelihood 0.9 and 0.2, respectively, whereas each of the six events will occur with likelihood 0.75. Assuming the events in each sequence to be independent of each other, the two-step sequence is more likely to occur (0.18 versus 0.178), yet only the six-step sequence will be seen as having the desirable downhill property. The overall plausibility of a scenario depends much more on the plausibility of its weakest links than on the number of links. I believe the author should respect this particular quirk of the readers, since it prevents him from resorting to facile but unlikely coincidences.

Even a downhill stream may have many twists and turns before it winds somewhere safe to sea. If it did not, observing its course would not provide much of an experience. The author is obliged, therefore, to provide the necessary surprises for readers and viewers, and obstacles for the characters, to keep audience interest alive. The repertoire of stratagems is huge, too huge to be surveyed or even to be classified. Some of them are closely linked with the genre. Within the theater, comedy, drama, and tragedy have different means at their disposal. Whereas comedy often relies on *misunderstandings* to generate tensions, drama and even more so tragedy may rely on *ignorance*. As misunderstandings are dissipated, felicity ensues; as ignorance is lifted,

[7] To be sure, redundancy is not always to be eschewed, since it can serve an aesthetic function. To convey boredom, redundancy may be more effective than a mere authorial statement. Yet even then, there would be a point when the repetition would bore the *reader* rather than evoking the boredom of the character.

disaster occurs. Novelists can add their own voices to those of the characters to generate uncertainty, as long as they do not deliberately mislead the readers.

I am now in a position to say what I mean by the "right interpretation" of a text. As I stated at the outset, this is a question of explanation. Since all explanations are causal (including those that cite intentions as causes) and since a cause must precede its effect, it follows that *actual* audience perceptions of the work are strictly irrelevant. Intended perceptions, by contrast, can be part of the explanation. Among the antecedent causes of the work, the authorial intention is not all that matters. The unconscious attitudes of the author may also influence it. Thus Jules Verne's *L'île mystérieuse* may have been shaped by his anti-racist intentions as well as by his racist prejudices. For the sake of brevity, however, I shall limit myself to conscious intentions.

An interpretation of a work of literature, then, is a claim that important features of the work can be traced back to decisions that the author made for the purpose of enhancing the aesthetic value of the experience that some specific audience could be expected to derive from the work. To make a claim of this kind, literary critics must proceed just as other scholars do. They can appeal to drafts, when they exist, and to statements by the author about the work, Stendhal's marginalia, for example. They can appeal to other works by the same author, to see whether a similar pattern of choices is observed. They can refer to contemporary works, to distinguish the conventions that frame choices from the choices themselves. They can draw on other contemporary sources to determine the audience expectations that may have constrained the author.

In doing all this, their method is in no way different from that of other historians. As other historians do, they face the problem that the data are essentially finite, because the past is not amenable to experiments. And as other historians do, they can try to minimize the temptations of "data mining" by triangulating old sources, looking for new sources, and drawing out novel implications of their interpretation to be tested against evidence (Chapter 3). They may differ from other historians in that their interpretation more often, although not invariably, goes together with *value judgments*. Did the author succeed, or approach closer to succeeding than to failing, in his or her aim of creating a local maximum of aesthetic value? Some writers, to be sure, do not have this aim. They may only be concerned with making money or writing propaganda, goals that have different rationality requirements. But if one can make out a plausible case for the hypothesis that the author had mainly aesthetic pretensions, it make sense to ask, as with any other aim, how well they were realized.

Earlier, I said that authorial failures may be intelligible. Authors, I have argued, are under a double pressure: they need to make the plot move on, and to do so through intelligible actions and statements by the characters. We may

blame them if they sacrifice the latter goal to the former – that is, if they sacrifice causality to teleology – but we can still *understand* why they do so. Even if causally implausible, Hamlet's procrastination could be made to seem teleologically intelligible in the light of Shakespeare's need to delay his vengeance until the end of the play. This, too, would be a piece of interpretation. Although obviously very different from an interpretation of the delay in terms of Hamlet's psychology and circumstances, it does answer the same question: why the delay? Although in a good work of literature everything can be explained twice over, imperfect works may only allow for one interpretation.

### Imputation of motives

In Chapter 4, I discussed how we sometimes impute motives to another agent by the hermeneutics of suspicion, assuming the worst of the motives that are compatible with the observed behavior. Literary critics, too, sometimes deploy this strategy. I shall discuss two examples: Paul Valéry's attempt at debunking Pascal and Stendhal, and some recent interpretations of Jane Austen's *Mansfield Park*.

Paul Valéry addressed one of the most famous sentences in the French language, Pascal's "Le silence éternel de ces espaces infinis m'effraie" ("The eternal silence of these infinite spaces fills me with dread"). In one comment, Valéry writes that "Every speech has several meanings, among which the most remarkable is surely the very cause of its being made ... To say: the eternal silence etc., is to state very clearly: I want to terrify you by my profundity and astonish you by my style." Elsewhere, he claims that "A distress that writes so well is not so complete that it hasn't salvaged some freedom of mind from the shipwreck, some sense of harmony, some show of logic or imagination, in contradiction to what the words themselves say ... If you want to attract or impress me, take care that I do not see your hand more distinctly than what it writes. I see Pascal's hand all too clearly." These statements undermine the very idea of literary or philosophical value, since any impressive achievement would serve as proof of an intention to impress.[8]

Valéry was seduced, against his will, by Stendhal's *Lucien Leuwen*. Rereading it, he says that "I was amazed to be [so deeply moved], because I could not and still can scarcely abide being deluded by a literary work to the point where I can no longer distinguish clearly between my own feelings and those

[8] Valéry's essay has been characterized as a "hostile philippic" of unusual virulence and malevolence. He asks whether Pascal "did not too deeply and bitterly resent the fame of Descartes," and suggests that if Pascal had not stipulated an opposition between salvation and knowledge, he might have inaugurated the infinitesimal calculus or non-Euclidean geometry.

suggested by the author's artifice. I see the pen and the person who is holding it. I do not care for, I have no need of, his emotions. I only ask him to let me into the secret of how it's done. But *Lucien Leuwen* brought about in me this miracle of a confusion which I abhor."

Valéry then engages in *backward reasoning*, from his feeling of being seduced to an intention to seduce. "I detect [in him] an element of calculation, a tendency to gamble on the reader of the future, a marked determination to attract by carelessness and apparent improvisation – which imply and suggest a 'just between you and me' relationship between the author and the unknown reader who is to be won over." He imputes to Stendhal the following recipe: "avoid the poetic style like the plague, and let the reader know you are avoiding it." He finds Stendhal's "accent three or four times too sincere; I sense a determination to be himself, to be genuine to the point of falsity." He repeats the charge he made against Pascal, that distortions are inevitable:

> How can we avoid selecting what is best out of the true we are working on? How can we avoid underlining, rounding off, touching up, adding color, trying to make it clearer, stronger, more disturbing, more intimate, more brutal than the model? In literature the true is inconceivable. Sometimes by simplicity, sometimes by oddity, sometimes by an exactness carried too far, sometimes by carelessness, sometimes by the confession of things that are more or less shameful, but always selected – as carefully selected as possible – always, and by every means in the author's power, whether he is Pascal, Diderot, Rousseau, or [Stendhal] ... We know very well that people only expose themselves for an effect.

Again, the fallacy is evident. Stendhal was a great admirer of Pascal, for reasons well expressed by another critic, who compared the maxims of Stendhal's *On Love* to those of La Rochefoucauld and La Bruyère "The arrows of our great moralists are more beautiful, and equally sound, but already stuck in their target; Stendhal's are seen in full flight. It is an impression that only Pascal conveys with greater force." Stendhal's novels, too, have a minimalist directness that helps to focus the reader's attention. Valéry's truism about the need for selection does not prove that Stendhal was writing to achieve an effect. Other writers – perhaps Rousseau is indeed an example – do illustrate the *temptation to write well*. As Pascal observed, "Those who construct antitheses by forcing the use of words are like those who put in false windows for the sake of symmetry." Stendhal's constant effort was to resist this temptation, to make the reader focus on the work, not on him, Stendhal.

Let me conclude by citing some recent examples of how interpretation may violate or ignore the demands of explanation. Several writers have claimed that Fanny Price in *Mansfield Park* is scheming and strategic, and that her seeming modesty is merely a stratagem deployed to win Edmund Bertram. They also argue, moreover, that her very name suggests "sex for money." These claims *fail two tests of intentionality*. First, there is no evidence in the novel for

imputing scheming intentions to Fanny Price. Although her modesty is in fact rewarded, that *consequence* of her behavior cannot explain it. Also, the hypothesis of a mercenary Fanny Price is refuted by her rejection of a marriage proposal from the better-situated Henry Crawford. Second, there is no evidence for imputing to Jane Austen an intention to make readers view Fanny Price as a semi-prostitute. Although the text may cause these associations to be produced in some modern readers, the writers in question offer no evidence that Austen intended her readers to associate "Fanny" with the heroine of the pornographic novel *Fanny Hill* or "Price" with payment for sex. These "interpretations by consequences" have much in common with functional explanations in the social sciences. They rely on arbitrary methods that are constrained not by facts but only by the limits of ingenuity of the scholars who propose them.

### Bibliographical note

The general approach I take in this chapter is often accused of embodying an "intentional fallacy." I agree with the responses of N. Carroll to this criticism, notably in "Art, intention and conversation," in G. Iseminger (ed.), *Intention and Interpretation* (Philadelphia: Temple University Press, 1992), and in "The intentional fallacy: defending myself," *Journal of Aesthetics and Art Criticism* 55 (1997), 305–9. In "Hermeneutics and the hypothetico-deductive method," in M. Martin and L McIntyre (eds.), *Readings in the Philosophy of the Social Sciences* (Cambridge, MA: MIT Press, 1994), D. Føllesdal offers an interpretation of *Peer Gynt* along similar lines, except that this play is not constrained by the convention that the events and characters that are described could have been real. I owe the observation that Hamlet's delay may have been due to dramaturgical concerns to E. Wagenknecht, "The perfect revenge – Hamlet's delay: a reconsideration," *College English* 10 (1949), 188–95. The comment on *Great Expectations* is from S. Rimmon-Kenan, *Narrative Fiction* (London: Methuen, 1983). The comment on God making Man in his own image is from a quirky and penetrating exploration of the analogy between divine and authorial rationality, D. Sayers, *The Mind of the Maker* (London: Methuen, 1941). I discuss the idea of works of art as local maxima in Chapter 3 of *Ulysses Unbound* (Cambridge University Press, 2000). That chapter also includes a fuller discussion of *Lucien Leuwen*. The idea of downhill versus uphill plots is inspired by D. Kahneman and A. Tversky, "The simulation heuristics," in D. Kahneman, P. Slovic, and A. Tversky (eds.), *Judgment Under Uncertainty* (Cambridge University Press, 1982). Paul Valéry's comments on Pascal and Stendhal are in his *Œuvres*, vol. I (Paris: Pléiade, 1957), pp. 458–71 and 553–82. Other relevant texts are cited in A. Rodriguez, *Paul Valéry et Pascal* (Paris: Nouvelles Éditions Debresse, 1977). A penetrating

criticism of Valéry's criticism of Pascal and Stendhal is J. Paulhan, *Paul Valéry, ou la littérature considérée comme un faux* (Paris: Gallimard, 1987). The comparison between Pascal and Stendhal is in J. Prévost, *La création chez Stendhal* (Paris: Mercure de France, 1971). The interpretations of *Mansfield Park* that I criticize are those of J. Heydt-Stevenson, "'Slipping into the ha-ha': bawdy humor and body politics in Jane Austen's novels," *Nineteenth-Century Literature* 55 (2000), 309–39, and of J. Davidson, *Hypocrisy and the Politics of Politeness* (Cambridge University Press, 2004).

*Part IV*

# Interaction

Social interaction can take many forms. (1) The outcome, for each agent, depends on the outcomes for others. This interdependence of outcomes can arise if the material or psychic welfare of others affects my own psychic welfare (Chapter 5). (2) The outcome of each can depend on the actions of all. This interdependence reflects general social causality (Chapter 17), illustrated in such phenomena as (human-made) global warming. (3) The action of each depends on the (anticipated) actions of all. This interdependence is the specific topic of *game theory* (Chapters 18 and 19), which also integrates (1) and (2) within its framework. (4) The beliefs of each depend on the actions of all. This interdependence can arise by a variety of mechanisms, such as "pluralistic ignorance" or "informational cascades" (Chapter 22). (5) The preferences of each depend on the actions of all. This interdependence is perhaps the least well-understood aspect of social interaction. Although I touch on some aspects of the question at various places, notably in Chapter 21, I offer no comprehensive account.

These interdependencies can arise through decentralized action by individuals who stand in no organized relation to each other (Chapter 23). Much of social life has more structure, however. Many outcomes occur through procedures of collective decision making – arguing, voting, and bargaining – through which groups of individuals reach decisions that are binding on them all (Chapter 24). Finally, institutions and constitutions create *rules* to put the incentives of individuals and goals of organizations in line with each other as well as *constraints* that have the dual effect of limiting and enabling the social agents (Chapter 25).

# 17 Unintended consequences

## Unintended consequences of individual behavior

Things do not always turn out the way we intend. Many events occur unintentionally. Sometimes, the causes are trivial, as when we press the accelerator instead of the brake or hit the "delete" button by mistake. Some mechanisms are more systematic, however. While there can hardly be a "general theory of unintended consequences," one can at least begin to compile a catalogue. I consider cases in which the consequences are not only unintended, but also unforeseen. Anticipated "side effects of action" are not intended for their own sake, especially if they are negative, but I shall not count them as "unintended consequences of action."

Unintended consequences can arise from individual behavior as well as from social interaction. Beginning with the former, we can use a simple extension of the desire–opportunity framework that was set out in Chapter 10 (see Figure 10.1).

While actions are shaped by desires (or preferences), they can also shape desires. Thus in addition to the intended outcome of an action, there is sometimes an unintended one: a change of desire. Addiction is a good example. Under the influence of addictive drugs, people begin to discount the future more heavily, thus weakening the deterrence effect of the long-term harm from addiction. Had this effect been anticipated, it might have prevented the agent from embarking on the "primrose path" to addiction, but typically it is not. Similar phenomena are observed in more ordinary situations. I go to the party intending to have only two drinks so that I can drive home, but after the second drink my resolve dissolves in alcohol and I take a third one. Had I known, I might have taken one drink only.

The "endowment effect," an implication of loss aversion (Chapter 14), also illustrates choice-induced but unintended preference change.[1] Many goods

[1] By accident, the term "endowment effect" has come to be used both for the tendency to overvalue items in one's possession and for the utility a person may derive in the present from utility in the past. The two meanings are entirely unrelated.

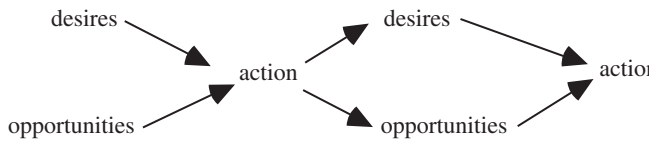



Figure 17.1

acquire greater subjective value for the owner than they had before she bought them, as shown by the fact that her minimal selling price typically exceeds her maximal buying price by a factor ranging from 2 to 4. Experiments show that prospective buyers underestimate the minimal resale price they would accept, showing that the preference change is indeed unforeseen. Another mechanism that could produce this "bolstering effect," the tendency, that is, to cast one's choices in a positive light once they have been made, is offered by the theory of cognitive dissonance (see Figure 17.1).

As an example of how action may shape *opportunities* in unintended and unforeseen ways, consider the bully who is able to get his way in transactions with others because they usually prefer to yield rather than stand up to him. An unintended consequence of his behavior may be that others shun him, so that he has fewer opportunities for transactions with them. He does well in each encounter, but he has fewer of them. The latter consequence may be not only unintended and unforeseen, but unperceived. As far as he can see, bullying works.[2] If he does notice the negative effects of his behavior, he might still persist in it if the positive effects outweigh them. In that case, the negative consequences will be foreseen but not intended for their own sake.

Often the choice of one option today will remove certain options from the feasible set in the future. This effect may be foreseen: my budget constraint may allow me to buy one car, but not two. Sometimes, however, the agent may not know that the choice has irreversible consequences. A peasant may have a piece of land on which there are some trees and some fields. To get more land for cultivation, and wood to burn, she cuts down the trees. The deforestation causes erosion, leaving her with less land for cultivation than she began with. In a set of cases that I shall discuss shortly, erosion may be the outcome of *collective* behavior, if for instance erosion occurs on the farmer's plot if and only if both she and her two neighbors carry out deforestation. But it is also possible and quite common for an individual single-handedly and unknowingly to undermine her future opportunities for action. The culprit is a

[2] This limited perspective is shared by some social scientists, who argue that emotions such as anger can be "rational," or at least adaptive, because they enable agents to get their way in encounters with others.

*cognitive deficit*: the agent cannot predict future consequences of present behavior. In other cases, it may occur through a *motivational deficit*: the agent attaches low weight to the (known and certain) future consequences compared to immediate gains (Chapter 6).

### Externalities

Let me now turn to unintended consequences of *interaction*, a theme that was one of the key ideas of the emerging social sciences, notably in the Scottish Enlightenment. In Adam Ferguson's memorable phrase, history is "the result of human action, but not the execution of any human design." His contemporary, Adam Smith, referred to an "invisible hand" that shapes human affairs. Half a century later, Hegel invoked the "cunning of reason" to explain the progress of freedom in history. About the same time, Tocqueville made a similar claim that in the progress of democracy, "everyone played a part: those who strove to ensure democracy's success as well as those who never dreamt of serving it; those who fought for it as well as those who declared themselves as its enemies." A few years later, Marx referred to people's "alienation" from their own action, claiming that "this fixation of social activity, this consolidation of what we ourselves produce as a material power above us, growing out of our control, thwarting our expectations, bringing to naught our calculations, is one of the chief factors in historical development up till now."

Among these writers only Adam Smith and Marx provided specific mechanisms for the production of unintended consequences. In modern language, they emphasized how *externalities* of behavior may aggregate to produce outcomes neither intended nor foreseen by the agents. In stylized form, imagine that each of many identical agents takes a certain action to promote his interest. As a by-product of that action, he also imposes a small cost or confers a small benefit (a negative or positive externality) on each of the other agents (and on himself). Each agent, then, is the target of many such actions. Adding up the effects, and then adding the sum to the private benefit of the agent caused by his action, we get the final outcome that the agents generate through their actions. Since we assume that they are identical, their initial states, the states they individually intend to bring about, and the states they collectively do bring about may each be represented by a single number, x, y, and z, respectively.[3]

[3] Many economists would not count all the phenomena I list here as externalities. They would include pollution, but not market-generated effects such as Keynesian unemployment. For my purposes, however, what matters is what they have in common: in pursuit of a benefit for himself, each individual imposes a small cost or benefit on everybody else *and on himself*. A firm laying off workers or cutting wages will cause a small reduction in the demand for its own

Suppose first that $z > y > x$, a positive externality. This was Adam Smith's main interest: when an agent directs his "industry in such a manner as its produce may be of the greatest value, he intends only his own gain, and he is in this, as in many other cases, led by an invisible hand to promote an end which was no part of his intention. Nor is it always the worse for society that it was no part of it. By pursuing his own interest he frequently promotes that of the society more effectually than when he really intends to promote it." In market competition the aim of each firm is to make a profit by producing more cheaply than the rivals, but in doing so they also benefit the customers. The customers, too, might in their capacity as workers or managers be in a similar position to benefit others through their competitive efforts. The result has been spectacular secular growth. The effect may or may not have been foreseen but was certainly "no part of" their intention.

Suppose next that $y > z > x$, a weak negative externality. The agents are made better off as a result of their effort, but, because of the costs they impose on each other, not as much as they expected to be. People commuting to work by car may be better off than they would be by using public transportation if the latter is poorly developed, but congestion or pollution prevents them from benefiting as much as they expected. If the externality is produced by congestion, they can hardly fail to notice it. If, however, it is produced by pollution, it might take a while before they understand that they are mutually harming themselves rather than being victims of (say) factory pollution.

Suppose finally that $y > x > z$, a strong negative externality. The agents are all made worse off as a result of everybody's trying to become better off. This was one of Marx's main charges against the decentralized capitalist economy. His main account of capitalist crises, the "theory of the falling rate of profit," had this general structure. To maintain or increase profits, he argued, each capitalist has an incentive to replace labor by machinery. When all capitalists do so simultaneously, however, they are collectively sawing off the branch they are sitting on, since the ultimate source of profit is the surplus value generated by labor. The argument is seductive but on closer analysis turns out to be wrong in all sorts of ways. More interesting is another observation that Marx made in passing and that later become a cornerstone of the theory of unemployment produced by John Maynard Keynes. Each capitalist, Marx noted, has an ambiguous relation to the workers. On the one hand, she wants the workers *she* employs to have low wages, since that makes for high profits. On the other hand, she wants all *other* workers to have high wages, since that makes for high demand for her products. Although it is possible for any one capitalist to have both of these desires satisfied, it is logically impossible for

products – but only a small one. If Henry Ford ever said "I want to pay my workers enough so they can afford to buy my cars," he was confused.

this to be the case for all capitalists simultaneously. This is a "contradiction of capitalism" that Keynes spelled out as follows. In a situation of falling profit, each capitalist responds by laying off workers, thus saving on the wage bill. Yet since the demand of workers directly or indirectly is what sustains the firms, the effect of all capitalists' simultaneously laying off workers will be a further reduction in profit, causing more lay-offs or bankruptcies.

There are many cases of this general kind. Overfishing, deforestation, and overgrazing ("the tragedy of the commons") may be individually rational, but collectively suboptimal or even disastrous. If each family in a developing country produces many children as insurance against poverty in old age, overpopulation will generate more poverty. In a water crisis, each individual who uses water for non-essential purposes causes a slight increase in the probability that the authorities may cut the water supply for a few hours each day, affecting essential purposes as well. These consequences may or may not be foreseen. A crucial feature of this category of unintended consequences is that even when they are foreseen, the behavior will be the same. As I explain in the next chapter, it is a *dominant strategy*: it is rational to choose it regardless of what others are doing.

### Internalities

A partially similar argument applies to "internalities," defined as the benefit or harm a person's choice at one time may confer on the welfare he derives from later choices. Metaphorically speaking, internalities are externalities that a person imposes on his "later selves." In the discussion of child custody summarized in Figure 13.3, I argued that time spent with the child creates a positive internality for the parent. Addiction provides an important example of negative internalities. The more a person has consumed in the past of an addictive substance, the less pleasure she derives from current consumption. This "tolerance" effect may also occur with non-addictive goods. Even if you love butter pecan ice cream, you are likely to be satiated if you have it five times a day. In addiction, however, past consumption has a further effect. While it makes current consumption less pleasurable than it would have been had the agent not consumed in the past, it also increases the welfare difference between consuming and not consuming in the present ("withdrawal"). Schematically, see Figure 17.2.

Thus regardless of whether she has abstained or consumed in the past, the agent is better off in the present by consuming than by not consuming. Consumption is a dominant strategy. At the same time, repeated consumption makes her worse off at all times (except for a few times in the beginning) than repeated abstention, just as the dominant strategy of having many children can make everybody worse off. There are of course obvious differences between

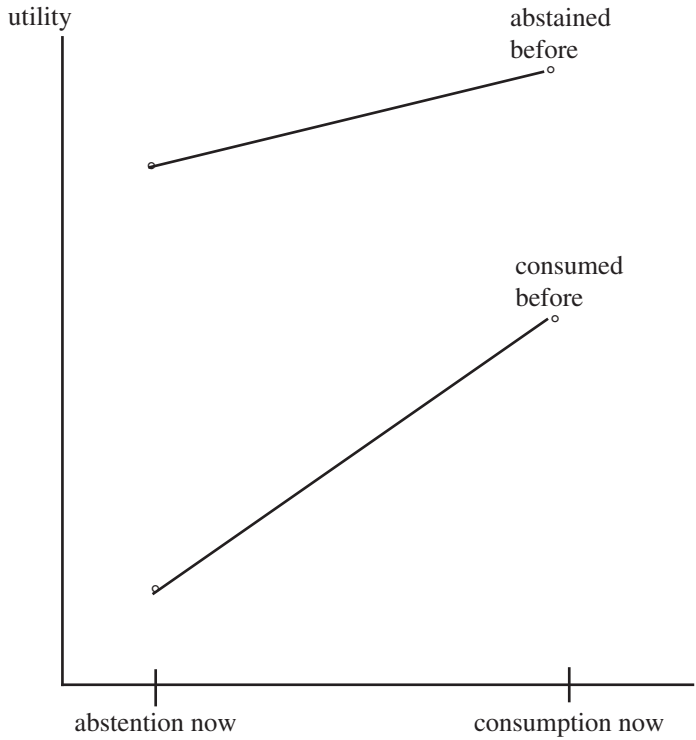



Figure 17.2

externalities and internalities. One is the temporal asymmetry: whereas all individuals can harm one another, later selves cannot hurt earlier selves.[4] Another is that the successive selves are really just time slices of *one* decision maker, whereas different individuals are not spatially distinct parts of one superorganism. Once *the* person (the one and only) understands that his present choices have a negative effect on the welfare he can derive from later choices, he has an incentive to change his behavior. Whether the incentive is strong enough depends on the severity of the withdrawal symptoms and on the extent to which the agent discounts future welfare. Some agents who would never have taken the first step had they known the consequences may choose not to quit once they are hooked.

[4] Nor benefit them: "Why should I do anything for future generations. They have never done anything for me" (Groucho Marx).

### The younger sibling syndrome

Unintended consequences of social action can also be produced by what I shall call the *younger sibling mechanism*. Before I explain this phrase, let me illustrate it with a famous example from economic theory, the "cobweb," also called the "hog cycle" because it was first put forward as an explanation of cyclical fluctuations in hog production. It has a much wider application, however. Fluctuations in the shipbuilding industry have often had the same pattern, with a seller's market followed by overinvestment and glut. When students make career choices on the basis of current demand for graduates, they may collectively undo the basis for their decisions.

Hog farmers must decide one year ahead of time how much they want to put on the market in the next year, a decision that is determined by the price they expect hogs to fetch and the cost of producing them. An increase in expected price will induce farmers to produce more, as reflected in the upward-sloping supply curve in Figure 17.3.

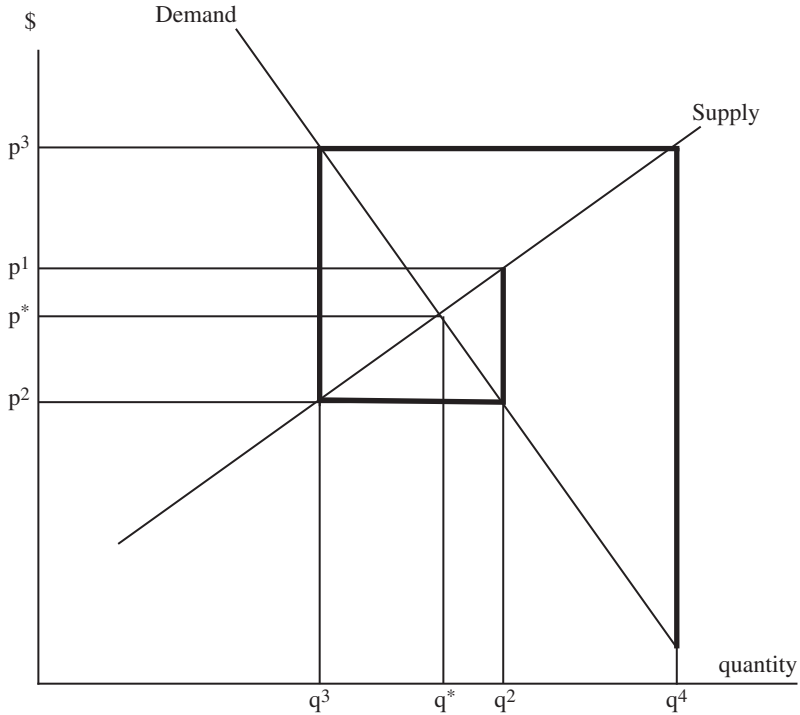



Figure 17.3

Assume that in year 1 the price for hogs is $p^1$. Expecting that prices will remain the same in year 2, farmers put $q^2$ on the market next year. At this volume, however, the market-clearing price is $p^2$ rather than $p^1$. Expecting that prices will remain at that level in year 3, farmers produce volume $q^3$ for that year. The market-clearing price will be $p^3$, inducing farmers to produce $q^4$ in year 4, and so on. In this case, prices and volumes form an outward spiral or "cobweb" pattern indicated by the bold lines in the diagram. Pleasant surprises alternate with unpleasant ones, but the expected outcome never occurs. If the relative slopes of the supply and demand curves are modified, the result could be an inward spiral converging to the equilibrium price p* and equilibrium volume q*.

There is something irrational about the behavior of the farmers. Each of them believes that *he* is free to vary his output to maximize his profits, while tacitly assuming that others are just mechanically doing what they did last year. While perhaps irrational, the behavior is certainly intelligible. A French philosopher, Maurice Merleau-Ponty, said that our spontaneous tendency is to view other people as "younger siblings."[5] We do not easily impute to others the same capacity for deliberation and reflection that introspection tells us that we possess ourselves, nor for that matter our inner turmoil, doubts, and anguishes (see Chapter 22). The idea of viewing others as being just as strategic and calculating as we are ourselves does not seem to come naturally.

Three examples from voting behavior can also illustrate the idea. Suppose I am a member of the left wing of the Socialist Party in my country. I would much prefer to have the Socialists rather than the Communists in power, but since the polls predict a solid Socialist majority I vote Communist to make my party move to the left. I do not, however, pause to ask myself whether other left-wingers might think along the same lines. If many of them do, the Communists might win. The intention to produce the top-ranked outcome (a Socialist victory with a strong Communist showing) may generate the third-ranked outcome (a Communist victory). In what is probably a more common scenario, if many voters stay home because they are confident that their party will win, it may lose. Finally, recall the Chirac example from Chapter 14. A possible explanation for his disastrous calling of early elections may have been his failure to anticipate that the voters would infer his beliefs from his decision rather than simply behave, mechanically, as they said in the polls that they would.

In some cases, agents may be able to learn from their mistakes and form approximately rational expectations (Chapter 6). For a case in which learning is irrelevant, consider a student who is deliberating over the choice between

[5] He actually wrote "younger brother."

law school and medical school. Assuming that she is motivated only by expected earnings, she will compare current incomes in the two professions to form her expectations about what she might earn three or six years hence. Once she has finished her studies in the chosen career, she may find that her income is substantially less than she expected, for the reasons underlying the cobweb model. Even if she understands where she went wrong, the understanding is irrelevant, since there is no repeat occasion on which she can use it.

The failure to see others as intentional and maximizing agents is observed when legislators or administrators propose policies that are undermined because agents adjust to them. According to Roman law, the stealing of a single horse or ox made a man a cattle thief, whereas it would not be a crime if he stole fewer than four pigs or ten sheep. A commentator on the law wrote that "in such a state of the law one would expect thefts of three pigs or eight sheep to become abnormally common." Today, this effect is observed in France, where firms with fewer than fifty employees are exempt from many burdensome administrative regulations. As a result, there are 24 percent fewer firms of between fifty and fifty-four employees than companies of between forty-five and forty-nine.[6] To promote security of employment, many countries have adopted legislation making it illegal to lay off workers who have been employed, say, two years or more. Employers rationally respond by outsourcing or by offering workers temporary contracts, thus reducing security of employment. Cities may build highways to reduce congestion, only to find that because more people take their cars to work the roads are just as jammed as before and more pollution is generated. The government may try to limit immigration to those who are married to a person who is already legally in the country, with the effect of inducing people to marry just for this purpose. Draft exemptions for students create an incentive to go to college. During the Vietnam War, dentists in Los Angeles charged between $1,000 and $2,000 to put braces on recruits who did not need them, because the law provided exemption for anyone under orthodontic care.

The younger sibling syndrome can have important social consequences, as some examples will bring out. Tocqueville notes that in the decades preceding the French Revolution, the upper classes publicly denounced the vices of the regime and its devastating impact on the people, as if the latter were deaf to what they were saying, "This reminds me of the sentiment of Mme Duchâtelet, who according to Voltaire's secretary did not mind undressing in front of her menservants, unpersuaded as she was that valets were men." The secretary, in his memoirs, wrote in fact that "great ladies regarded their lackeys only as

[6] French firms are subject to fifteen such numerical thresholds, each of which generates obligations for the employers. With twenty or more employees, a firm has to hire at least 6 percent disabled workers; with two hundred or more it has to provide a room for the trade union, and so on.

automata." This simultaneous show of contempt toward the lower classes and denunciation of their misery prepared minds for the Revolution. Similarly, to explain the Virginia slave rebellion of 1800, to which I referred in Chapter 10, Federalists cited the fact that the doctrine of liberty and equality had "been most imprudently propagated for several years at our tables while our servants were standing behind our chairs." More recently, the argument behind the "Phillips curve," according to which the government can choose, if it so desires, to realize low unemployment at the cost of high inflation, presupposes that the social actors are unaware of this policy. When governments tried to achieve this end, however, strategic behavior by rational trade unions and other actors undermined their efforts and produced instead "stagflation" – high inflation *and* high unemployment.

The Vietnam War can also be used to illustrate the younger sibling syndrome. On several occasions, American decision makers failed to see that both the South Vietnamese and the North Vietnamese governments would adapt strategically to their choices. The Americans sent ground troops to the South on the assumption that their ally would maintain its forces at an unchanged level. Instead, the government reduced its troops, leaving the *sum* of American and South Vietnamese troops more or less unchanged. (The US ambassador to Saigon warned against this possibility, but was overridden by the generals.) Similarly, the US bombed oil installations in North Vietnam on the assumption that the bombing would bring the country to its knees. Instead, China and the Soviet Union made up for the losses.[7]

Unlike the unintended consequences produced by externalities, those generated by the younger sibling mechanism may end when the agents understand it. There are no dominant strategies in the cases I have described, only strategies that are optimal on the (usually implicit) assumption that others are less rational than one is. Once all agents view each other as rational, their behavior may converge to a fully predictable outcome. All hog farmers will expect the equilibrium price to prevail. Acting on that expectation, they will produce the equilibrium volume. Their shared belief is self-fulfilling. This idea is the topic of the next chapter.

### Bibliographical note

The impact of addiction on time discounting is documented in L. Gordano *et al*., "Mild opioid deprivation increases the degree that opioid-dependent outpatients discount delayed heroin and money," *Psychopharmacology* 63 (2002), 174–82. The note offering a model of Andersen's tale is inspired by

[7] The assumption rested on Walt Rostow's analogy with the bombing of Germany at the end of World War II, ignoring the fact that the Germans had no other source to which they could turn.

C. C. von Weiszäcker, "Notes on endogenous change of tastes," *Journal of Economic Theory* 3 (1971), 345–72. For a discussion of Marx on unintended consequences, see my *Making Sense of Marx* (Cambridge University Press, 1985), Chapter 1.3.2 and *passim*. The addiction model derives from G. Becker and K. Murphy, "A theory of rational addiction," *Journal of Political Economy* 96 (1988), 675–700. A superb conceptual study of unintended consequences is T. Schelling, *Micromotives and Macrobehavior* (New York: Norton, 1978). For the idea of internalities, see R. Herrnstein *et al*., "Utility maximization and melioration: internalities in individual choice," *Journal of Behavioral Decision Making* 6 (1993), 149–85.

# 18 Strategic interaction

## Strategic interaction with simultaneous choices

The invention of *game theory* may come to be seen as the most important single advance of the social sciences in the twentieth century. The value of the theory is partly explanatory, but mainly conceptual. In some cases it allows us to explain behavior that previously appeared as puzzling. More important, it illuminates the structure of social interaction. Once you see the world through the lenses of game theory – or "the theory of interdependent decisions," as it might better be called – nothing looks quite the same again.

I first consider games in which agents make simultaneous decisions. The goal is to understand whether and how n agents or *players* may achieve an unenforced coordination of their *strategies*. Often, we shall look at the special case of n = 2. A strategy can be the simple choice of an *action*, as when a poker player decides to bluff after looking at his hand. It can also be the choice of a *rule*, as when he decides to bluff only when he is dealt the seven of hearts. Finally, a player can use a *mixed strategy*, that is, assign probabilities to the possible actions and choose one of them with the assigned probability. In shooting a penalty, a soccer player might, for instance, mentally flip a coin between aiming at the right or the left side of the goal. The goal keeper might use the same procedure to decide whether to go to the right or the left.

The players may be able to communicate with each other, but not to enter into binding agreements. (They can make promises, but the decision whether to *keep* the promise will only recreate the game.) To any n-tuple of strategies, one chosen by each agent, there corresponds an *outcome*. Each agent ranks the possible outcomes according to his or her *preference order*. When needed, we shall assume that the conditions for representing preferences as cardinal utilities are satisfied (Chapter 13). The *reward structure* is the function that to any n-tuple of strategies assigns an n-tuple of utilities. Although the word "reward" may suggest a monetary outcome, the word will be used to refer to psychological outcomes (utilities and ultimately preferences). When the monetary or material reward structure and the psychological reward structure diverge, only the latter is relevant.

As briefly mentioned in the last chapter, an agent may have a strategy that is *dominant* in the sense that regardless of what others do, it yields a better outcome for her than what she would get if she chose any other strategy. Her *outcome* may depend on what others do, but her *choice* does not. In other cases, there is genuine interdependence of choices. If others drive on the left side of the road, my best response is to drive left too; if they drive on the right, my best response is to drive right.

An *equilibrium* is an n-tuple of strategies with the property that no player can, by deviating from his equilibrium strategy, unilaterally bring about an outcome that he strictly prefers to the equilibrium outcome. Equivalently, in equilibrium the strategy chosen by each player is a best response to the strategies chosen by the others, in the weak sense that he can do *no better* than choosing his equilibrium strategy if others choose theirs. The strategy need not, however, be optimal in the strong sense that he would do *worse* by deviating unilaterally. In the general case, a game may have several equilibria. We shall see some examples shortly. Assume, however, that there is only one equilibrium. Assume moreover that the reward structure and the rationality of all players are common knowledge.[1] Under these assumptions, we can predict that all agents will choose their equilibrium strategy, since it is the only one that is based on rational beliefs about what others will do.

Some games with a unique equilibrium turn upon the existence of dominant strategies. The phrase "turn upon the existence of dominant strategies" can mean one of two things, illustrated in panels A and B of Figure 18.1.[2] In an accident involving two cars, both are harmed. In an accident involving a pedestrian and a car, only the former is harmed. Car–car accidents occur if at least one driver is careless. If both are careless, the outcome is worse. Car–pedestrian accidents occur only if both are careless. Taking due care is costly. From these premises, it follows that in the car–car case, taking care is the dominant strategy for each driver. In the car–pedestrian case, no care is dominant for the driver. The pedestrian has no dominant strategy, since due care is the best response to no care and no care the best response to due care.

[1] A fact is common knowledge if all know it, all know that all others know it, all know that all others know that all others know it, and so on. To avoid reliance on the phrase "and so on," which suggests an infinite sequence of beliefs, the idea may also be stated as follows: there is no n such that the fact is common knowledge up to level n in the sequence but not at level n + 1. For a simple illustration, common knowledge may be realized in a classroom. When the teacher tells a fact to the students, they all know it, know that others know it, and so on.

[2] By convention, the first number in each cell represents the pay-off for the "row player" who chooses between the top and bottom strategies, and the second the pay-off for the "column player" who chooses between the left and right strategies. Depending on the context, the pay-offs may be cardinal utilities, ordinal utilities, money, or anything else that the players may be assumed to maximize. In Figure 18.1, pay-offs may be seen as standing for ordinal utilities, reflecting preferences over outcomes. Here and later, equilibria are circled.

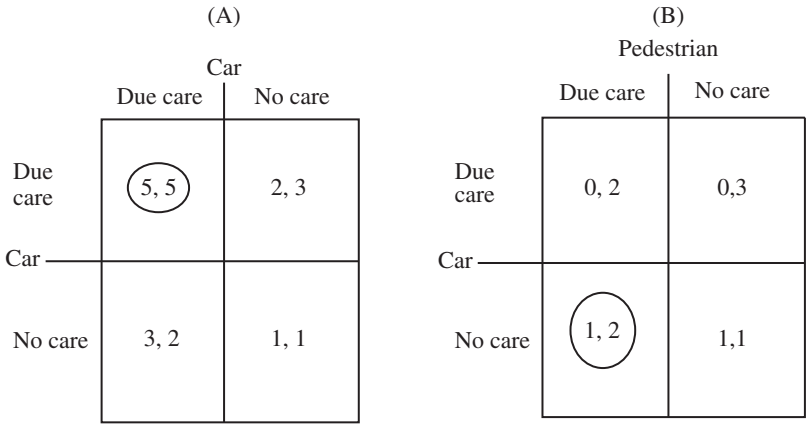



Figure 18.1

Since he knows that the driver has no care as a dominant strategy and, being rational, will choose it, the pedestrian will nevertheless choose due care.[3]

Games in which all players have dominant strategies are quite common and empirically important, as we shall see. Theoretically, they are somewhat trivial, except when they are repeated over time. Games in which some players have dominant strategies that can induce clear-cut choices in others are less common but also important. They have stronger informational requirements, however, since in our example the pedestrian needs to know the possible outcomes for the driver as well as for himself, whereas the two drivers only need to know their own outcomes. Often, we can impute dominant strategies to others without much trouble. We do not usually, for instance, look both ways before crossing a one-way street because we assume that the fear of drivers of being liable for an accident will make them obey the one-way rule.

A special class of games has *coordination equilibria*, often called "conventions," in which each player not only has no incentive to deviate unilaterally, but also would prefer that nobody else does so. In an equilibrium in which everybody drives on the right side of the road, an accident might occur if I deviate *or* if anyone else does. In this case, the equilibrium is not unique, since driving on the left side has the same properties.[4] Often it does not

[3] According to some legal analyses, an important function of tort law is to use the system of fines and damages to change the reward matrix so that the emerging equilibrium has some desirable property (efficiency or fairness).

[4] Although the non-uniqueness does not follow from the formal definition, this seems to be a general feature of real-life coordination games.

matter what we do as long as we all do the same thing. The meanings of words are arbitrary, but once they are fixed, they become conventions. In other cases, it does matter what we do, but it is more important that we all do the same thing. I return to some examples shortly.

### Two duopoly examples

Some games have unique equilibria that do not turn upon the existence of dominant strategies. Duopoly behavior is an example (see Figure 18.2). When two firms dominate a market, lower production by one firm will induce higher prices and an expansion of production by the other firm. In other words, each firm has a "best response" schedule that tells it how much to produce as a function of the output of the other firm. In equilibrium, the output of each firm is a best response to the output of the other. This statement does not imply that they could not do better. If they formed a cartel and restricted their production to below-equilibrium levels, both would earn greater profits. Yet these collectively optimal levels of production are not best responses to each other. The firms are, in fact, facing a Prisoner's Dilemma (defined in Figure 18.2).

For another case of duopoly, consider two ice cream vendors on a beach, trying to find the best location for their stalls on the assumption that customers (assumed to be evenly distributed across the shoreline) will go to the closer

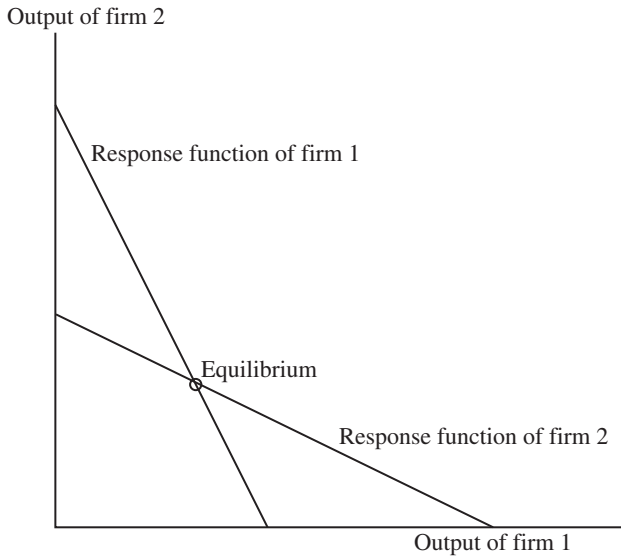



Figure 18.2

stall. There is no dominant strategy. If one of them puts up a stall some distance left of the middle of the beach, the best response of the other is to position himself immediately to the right, to which the best response of the first is to move right again, and so on, until their stalls are beside each other at the middle of the beach. This unique equilibrium is obviously not the best for the customers in the aggregate. For them, the best outcome is one in which each stall is positioned halfway between the middle and one end of the beach. Although this outcome is just as good for the sellers as the equilibrium outcome, these positions are not best responses to each other. This model has also been applied to explain the tendency for political parties (in a two-party system) to move toward the middle of the political spectrum.

Suppose, however, that when both stalls are at the middle customers close to the ends abstain from buying ice cream because it would melt by the time they walked back. If no customer is willing to walk more than half the length of the beach, one-quarter to get to the stall and one-quarter to get back, the optimal consumer outcome is also the unique equilibrium since neither has an incentive to relocate. Suppose the beach is 1,000 meters long. If the seller at 750 meters moves his stall to 700 meters, he will lose the fifty customers between 950 and 1,000 meters who are not willing to walk more than 500 meters and gain the twenty-five customers between 475 and 500 meters to whom his stall is now closer than the other – a net loss. A similar argument might also explain why political parties never converge fully to the middle, since extremists at either end might prefer to abstain rather than vote for a centrist party. In addition, as I noted at the end of Chapter 11, it is simply not plausible to view vote maximization as the only aim of political parties.

### Some frequently occurring games

A few simple interaction structures, with pay-offs as in Figure 18.3, occur very often in a great variety of contexts. C and D stand for "cooperation" and "defection." In the Telephone Game the column player is the one who first called. In the Focal Point Game, A and B can be any pair of actions such that both players would prefer to coordinate on either than not to coordinate but are indifferent between the two ways of coordinating.

The games illuminate the structure of the two central issues of social interaction – *cooperation* and *coordination*. In a society with no cooperation for mutual benefit, life would be "solitary, poor, nasty, brutish, and short" (Hobbes). That it would be *predictably* bad is a meager consolation. In a society where people were unable to coordinate their behavior, unintended consequences would abound and life would be like "a tale told by an idiot, full of sound and fury, signifying nothing" (*Macbeth*). Both cooperation and

| | C | D |
|---|---|---|
| C | 3, 3 | 0, 4 |
| D | 4, 0 | 1.5, 1.5 |

Prisoner's Dilemma

| | C | D |
|---|---|---|
| C | 4, 4 | 1, 3 |
| D | 3, 1 | 2, 2 |

Stag Hunt/ Assurance Game

| | C | D |
|---|---|---|
| C | 2, 2 | 1, 3 |
| D | 3, 1 | 0, 0 |

Chicken

| | Ballet | Boxing |
|---|---|---|
| Ballet | 1, 2 | 0, 0 |
| Boxing | 0, 0 | 2, 1 |

Battle of the Sexes

| | Redial | Do not redial |
|---|---|---|
| Do not redial | 2, 2 | 0, 0 |
| Redial | 0, 01, 1 | |

Telephone Game

| | A | B |
|---|---|---|
| A | 1, 1 | 0, 0 |
| B | 0, 0 | 1, 1 |

Focal Point Game

Figure 18.3

coordination sometimes succeed, but often fail abysmally. Game theory can illuminate the successes as well as the failures.

The Prisoner's Dilemma (PD), the Stag Hunt, and Chicken involve in one way or another the choice between cooperation and defection (non-cooperation). The Prisoner's Dilemma is so called because the following story was used to illustrate it in an early discussion. Each of two prisoners, who have been involved in the same crime but are now in separate cells, is told that if he informs on the other but she does not inform on him, he will go free and she will go to prison for ten years; if neither informs on the other, both will go to prison for one year; and if both inform on each other, both will go to prison for five years.[5] Under these circumstances, informing is a dominant strategy, although both would be better off if neither informed. The outcome is generated by a combination of the "free-rider temptation" (going free) and the "fear of being suckered" (getting ten years).

The negative externalities discussed in the last chapter can also be viewed as many-person PDs. Some other examples follow. For each worker (assuming selfish motivations) it is better to be non-unionized than to join a union, even when it is better for all if all join and gain higher pay. For each firm in a cartel it is better to break out and produce a high volume to exploit the high prices caused by the output restrictions of the other firms, but when all do that, prices fall to the competitive level; profit maximization by each firm undermines the maximization of joint profits. The Organization of Petroleum Exporting Countries (OPEC) cartel is vulnerable in the same way. Other examples are situations in which everybody has to run as fast as he can to stay in the same place, such as the arms race between the United States and the former Soviet Union, political advertising, or students writing papers for a teacher who "grades on the curve." I offer many other examples in Chapter 23.

The idea of the Stag Hunt is often imputed to Jean-Jacques Rousseau, although his language was somewhat opaque.[6] In more stylized form, it involves two hunters who can choose between hunting a stag (C) or a hare (D). Each can catch a hare by himself, but the joint effort of both is necessary (and sufficient) to catch a stag. Half a stag is worth more than a hare. It takes more time and effort to catch hares when both are trying because the noises the hunters make scare them away. As in the Prisoner's Dilemma, there is a risk of

[5] The pay-offs for the Prisoner's Dilemma in Figure 18.3 might seem artificial. For the present purposes, all that matters is the (ordinal) ranking of the outcomes. Later, the pay-offs will be reinterpreted as monetary rewards.

[6] "If a deer was to be taken, every one saw that, in order to succeed, he must abide faithfully by his post: but if a hare happened to come within the reach of any one of them, it is not to be doubted that he pursued it without scruple, and, having seized his prey, cared very little, if by so doing he caused his companions to miss theirs." This could be read as saying that pursuing hares is a dominant strategy.

being a sucker, hunting a stag while the other goes for a hare. There is no free-rider temptation, however. The game has two equilibria, in the upper left-hand and lower right-hand cells.

Although the first equilibrium is clearly better, it may not be realized. To see why this might happen we can drop the assumption that the pay-off structure is common knowledge and allow the agents to have mistaken beliefs about the pay-off structure of other agents. Actions taken on these beliefs will form an *equilibrium in a weak sense* if, for each agent, the actions taken by the others confirm his beliefs about them. Assume, for instance, that in a Stag Hunt each agent falsely believes the others to have PD preferences. Given that belief, the rational action is to defect, thus confirming the belief of the others that *he* has PD preferences. This society might end up with high levels of tax evasion and corruption. I return to such cases of "pluralistic ignorance" in Chapter 22. In another society, where people correctly believe others to have Stag Hunt preferences, a good equilibrium will emerge in which people pay their taxes and do not offer or take bribes. "Cultures of corruption" might be a belief-dependent, not a motivation-dependent, phenomenon.

International control of infectious diseases can have the structure of a Stag Hunt. If only one country fails to take the appropriate measures, others will not be able to protect themselves.[7] For another example, consider counterterrorist measures. If only one of two nations invests in such measures, it benefits the other as well as itself. If the costs exceed the benefits to itself, it will not invest unilaterally. Yet if both invest, the ability to pool information may lead to a greater security level for each than it could achieve by exploiting the investment of the other.

In these examples, the pay-off structure arises from the causal nature of the situation. In the Stag Hunt and the disease control case, the "threshold technology" implies that individual efforts are pointless. In the counterterrorism case, the underlying cause is something like economies of scale: ten units of effort have more than twice the effect of five units. In other cases, the pay-off structure is due to the fact that the agents care for other things than their own material rewards. In such cases, it is more common to refer to the game as an Assurance Game (AG). Even if the material pay-off structure is that of a PD, each individual may be willing to cooperate if he is *assured* that others will. The desire to be fair, or the reluctance to be a free rider, may overcome the temptation to exploit the cooperation of others. Alternatively, altruistic preferences may transform a PD into an AG.

Let us interpret the pay-offs in the PD in Figure 18.3 as monetary rewards and assume that each person's utility equals his monetary reward plus half the

[7] This is a huge simplification, made simply for the sake of illustration.

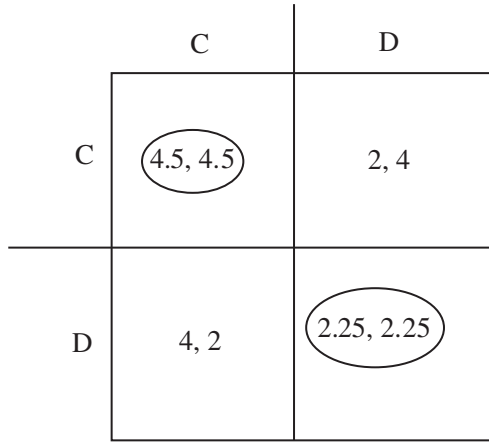



Figure 18.4

monetary reward of the other. In that case, the utility pay-off will be as in Figure 18.4 – an AG. The PD may also be transformed into an AG by a third mechanism, if an outside party attaches a penalty to the choice of the non-cooperative strategy D. If we again interpret the pay-offs in the PD in Figure 18.3 as monetary rewards, *and* assume that this is all the agents care about, deducting 1.25 from the reward to defection will turn it into an AG. A labor union might, for instance, impose formal or informal sanctions on non-unionized workers. Finally, one might transform a PD into an AG by rewarding cooperation, for example, by offering a bonus or bribe of 1.25 to cooperators. Promises of reward have to be respected, however, whereas a threat does not have to be carried out if it works. If the free-rider pay-off is very high, the benefits from cooperation may not be large enough to fund the bribes.[8] In some cases, though, rewards are used. Workers who join a union may benefit not only from higher wages, which sometimes accrue equally to non-unionized workers, but also from pension plans and cheap vacations offered only to members. Again, I refer to Chapter 23 for details and examples.

The game of Chicken is named after a teenage ritual from the 1955 movie *Rebel Without a Cause*. Los Angeles teenagers drive stolen cars to a cliff and play a game in which two boys simultaneously drive their cars off the edge of the cliff, stopping at the last possible moment. The boy who stops first is "chicken" and loses. In another variant, two cars drive toward each other and

[8] Whether one uses punishments or rewards, the costs of establishing the system and monitoring the agents also have to be funded by the gains from cooperation. In practice, this can easily make such arrangements impossible or wasteful.

the one who swerves first is "chicken." In each of the two equilibria, each agent does the opposite of the other. Even with common knowledge of the pay-off structure and of the rationality of the agent, we cannot predict which of the equilibria (if any) will be chosen. From the point of view of rational choice, the situation is *indeterminate* (see Chapter 13). In the second ("swerve") version of the game, a player might try to break the indeterminacy by (visibly) blindfolding himself, thus inducing the other to swerve. Yet this creates the same predicament with the two options being "blindfolding" and "not blindfolding" rather than "swerving" and "not swerving."[9] It is a deeply frustrating situation.

On one understanding of the arms race, it has the structure of Chicken. The Cuban missile crisis is often cited as a case in which the two superpowers were locked in a Chicken-like confrontation and the USSR "blinked first." Another example is that of two farmers who use the same irrigation system for their fields. The system can be adequately maintained by one person, but both farmers gain equal benefit from it. If one farmer does not do his share of maintenance, it may still be in the other farmer's interest to do so. The "Kitty Genovese" case can also be seen in this perspective, if we assume that each neighbor would prefer to intervene if and only if nobody else did.

Turning now to questions of *coordination*, consider first the Battle of the Sexes. The stereotype behind the story is the following. A man and his wife want to go out for the evening. They have decided to go either to a ballet or to a boxing match after work and to settle the final choice over the telephone. His phone breaks down, however, so they have to decide by tacit coordination. They have a common interest in being together, but divergent interests about where to go. As does the game of Chicken, this game has two equilibria, coordinating on the ballet or on the boxing match. And as in that game, there is no way common knowledge of the pay-off structure and of rationality will tell the couple where to meet. Once again, the situation is indeterminate.

Games of this kind arise when coordination can take many forms, all of which are better for all agents than no coordination at all, but each of which is preferred by some agents to the others.[10] In social and political life, this seems to be the rule rather than the exception. All citizens may prefer any political constitution (within a certain range of possible regimes) to no constitution at

[9] Similarly, the "solution" to the Prisoner's Dilemma that consists of each person's promising to cooperate merely recreates the PD with the choices being "keeping the promise" and "reneging."

[10] As we shall see later (Chapter 24), this question of dividing the benefits from cooperation can also be studied within *bargaining theory*, a specialized branch of game theory.

all, because long-term stability is important in enabling them to plan ahead. When the law is fixed and hard to change, one can regulate one's behavior according to it. Yet each interest group may prefer a specific constitution in the range over the others: creditors lobby for a ban on paper money in the constitution, each political party favors the electoral system that will favor it, those with a strong candidate for the presidency want that office to be strong, and so on (see Chapter 25).

Multiple coordination equilibria also arise when different societies initially develop different standards of weight, length, or volume and later discover the potential benefits of a common solution. Continental Europe and the Anglo-Saxon world retain separate standards in these areas. Unlike the case of multiple constitutional solutions, the obstacle to agreement is not permanent divergence of interest, but short-term transition costs. The choice of standard might also, however, be a game of Chicken. Assume, implausibly, that the standard is written into the constitution as an entrenched clause (immune to amendment). Each country will then have an incentive to commit itself before the other does.

The Telephone Game is defined by the need for a rule to tell the parties what to do when a phone conversation is accidentally interrupted. There are two coordination equilibria: the redialing is done by the person who made the call in the first place or by the person who received it. Either rule is better than having both redial or neither. Yet in this case, unlike the Battle of the Sexes, one equilibrium is better for both than the other. It is more efficient to have the caller do the redialing, since he is more likely to know which number to call. Rational, fully informed agents will converge on the superior coordination equilibrium. This statement ignores, however, the cost of redialing. If the cost is large, the game becomes a Battle of the Sexes.

Consider finally the Focal Point Game, which can be illustrated by a variant of the Battle of the Sexes. The spouses have agreed to watch a movie that is playing both in movie theater A and movie theater B but have postponed the choice of venue. We assume that neither is closer or otherwise more convenient than the other. As in the Battle of the Sexes, information, rationality, and common knowledge by themselves will not tell them where to go. There might, however, be a psychological cue in the situation that will serve as a "focal point" for coordination. If the couple had their first date in theater A, this might make them converge to that location. In this case, the cue is a purely private event. In other cases, cues might be shared by a large population. Among New Yorkers, for instance, folklore says that if you get separated from your companion you meet at noon under the main clock at Grand Central Station. And even when there is no folklore, many people would still go to the railway station, since in many cities the railway station is the most important

building of which there is only one.[11] Its uniqueness renders it attractive as a focal point. Noontime has the same property.[12]

This focal point effect is easily demonstrated in experiments. If you ask all members of a group to write down a positive integer (whole number) on a piece of paper and tell them that they will get a reward if all write down the same number, they invariably converge on 1. There is a unique smallest integer, but no unique largest one. In other contexts, 0 may emerge as the unique focal point. In debates during the Cold War whether the United States might use tactical nuclear weapons without triggering an escalation into full-blown nuclear war, various ideas were suggested for a "bright line" that would allow limited use. In the end, it was decided that *no use* was the only focal point.

Pascal made a similar observation about the importance of custom: "Why do we follow old laws and old opinions? Because they are better? No, but they are unique, and remove the sources of diversity." Elsewhere he wrote

> The most unreasonable things in the world become the most reasonable because men are so unbalanced. What could be less reasonable than to choose as ruler of a state the oldest son of a queen? We do not choose as captain of a ship the most highly born of those aboard. Such a law would be ridiculous and unjust, but because men are, and always will be, as they are, it becomes reasonable and just, for who else could be chosen? The most virtuous and able man? That sets us straight away at daggers drawn, with everyone claiming to be most virtuous and able. Let us then attach this qualification to something incontrovertible. He is the king's eldest son: that is quite clear, there is no argument about it. Reason cannot do any better, because civil war is the greatest of evils.

Other countries, from the Roman Empire onward, were exposed to constant civil wars because of the absence of a rule of succession. In the choice of king in the French Restoration, Talleyrand successfully argued that the legitimate heir of the last king of France was the unique focal point that could prevent divisive conflicts. As he wrote in his memoirs, "An *imposed* King would be the result of force or intrigue; either would be insufficient. To establish a durable system that will be accepted without opposition, one must act on a principle." Later, Marx argued that the Republic of 1848 owed its existence to the fact that it was the second-best option for each of the two branches of the royal family. Tocqueville made a similar observation to explain the stability of the rule of Napoleon III. Democracy, too, can be seen as a focal-point solution. When there are many competing qualitative grounds on which people can claim

[11] In New York City, those ignorant of the folklore would not go to Grand Central Station, since the presence of Penn Station makes it non-unique. Instead, they might coordinate on the Empire State Building.

[12] Although midnight, too, is a focal point, it is inferior to noontime because of the inconvenience.

superiority – wisdom, wealth, virtue, birth – the quantitative solution of majority rule acquires unique salience. Former colonial countries in which tribes speak different languages may choose the language of the colonizer for official purposes. Litigating parties easily converge on a proposal that is everybody's second-best option.

In June 1989, the reburial of Imre Nagy provided a focal point for 250,000 people to march in the streets of Budapest to signal their disaffection with the regime. To *call* for a demonstration would have been dangerous for the organizers, but the spontaneous convergence on Heroes' Square needed no coordinator. In conflict situations, focal points can have quite different effects. In the Crimean War the French General Pélissier decided to stage the second attack on Sebastopol on June 18, 1855, because he wanted to please Napoleon III by gaining a victory on the anniversary of the Battle of Waterloo. As this date and its importance to the French were common knowledge, the Russians were able to anticipate and defeat him.

One lesson from this survey is that a given real-world situation can be modeled as several different games, depending on additional assumptions. The arms race has been modeled as a PD, as Chicken, and as an AG. Joining the labor union may be a PD or an AG. Redialing has been seen as a Battle of the Sexes or as a Telephone Game. Coordination of weights and measures could be a game of Chicken or Battle of the Sexes. The fine grain of interaction structures may not be immediately visible. By forcing us to be explicit about the nature of the interaction, game theory can reveal unsuspected subtleties or perversities.

### Sequential games

Let me turn more briefly to games in which agents make *sequential decisions* (I discuss such games at greater length in the next chapter) and begin by a simple example that demonstrates the power of game theory to clarify interaction structures that were only dimly understood earlier.[13]

In Figure 18.5, two armies are confronting each other at the border of their countries. General I can either retreat, leaving the status quo (3, 3) in place, or invade. If he invades, General II can either fight, with outcome (2, 1), or concede a contested piece of territory with outcome (4, 2). Before General I makes his decision, General II may be able to communicate an intention to fight if attacked, hoping to induce General I to choose (3, 3) rather than (2, 1). However, *this threat is not credible*. General I knows that once he invades, it will be in General II's interest to concede rather than to fight. The unique

[13] I retain the assumption that rationality and information are common knowledge.

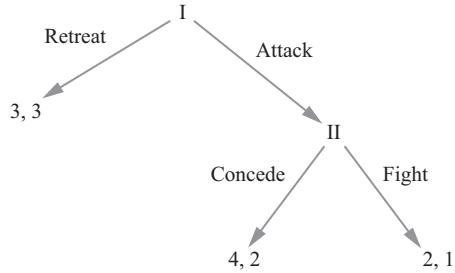



Figure 18.5

equilibrium outcome is (4, 2). This equilibrium concept is not the static "best-response" concept we have been discussing so far. Rather, it is a dynamic concept that begins with the later stages of the game and works back to the earlier ones. (The technical term is "backward induction.") First, we ask what it would be rational for General II to do if General I invaded. The answer, "Concede," leads to the outcome (4, 2). General I's choice, therefore, is between a course of action leading to (3, 3) and one leading to (4, 2). Being rational, he chooses the latter.

Promises as well as threats need to be credible of they are to affect behavior. Allowing for communication in the Trust Game (Chapter 20), for instance, the trustee might try to induce the investor to make a large transfer by promising to make a large back transfer. If there is nothing to hold him to his word, the promise is not credible. Economic reform in China was vulnerable to this problem. When the government introduced market reforms in agriculture in the 1980s, it promised the farmers fifteen-year leases on the land to give them an incentive to improve it. Since there is no way of holding an autocratic government to its promise, many farmers disbelieved it and used the profits for consumption instead. An autocratic government is *unable to make itself unable* to interfere (see chapter 24).

The notion of credibility is central in the "second-generation" game theory that began around 1975. (The first generation began around 1945.) Once we take the idea seriously, we are led to ask how agents might *invest in credibility* to lend efficacy to their threats and promises. There are several mechanisms. One is by *reputation-building*, for instance by investing in a reputation for being somewhat or occasionally irrational. Thus President Nixon, encouraged by Henry Kissinger, deliberately cultivated an erratic style to make the Soviets believe he might act against the American interest if they provoked him. Also, people might carry out threats when it is not in their interest to do so in order to build a reputation for toughness that will make others believe their threats on later occasions. I return to these questions in Chapter 24.

Another mechanism is *precommitment*, discussed in Chapter 14 and Chapter 15. There, precommitment was viewed as a second-best rational response to the agent's proclivity to behave irrationally. In the strategic context, precommitment can be fully rational. In the game depicted in Figure 18.5, General II might build a "Doomsday machine" that would automatically launch a nuclear attack on the other country in the case of invasion. If both the existence of this machine and the fact that its operation is outside the control of country II were common knowledge, it would deter the invasion. Alternatively, General II might use the strategy of "burning his bridges," that is, of cutting off any possibility of retreat. If General II has no option but to fight if attacked, General I will obtain a payoff of 2 if he attacks – less than he would obtain by retreating. Less can be more, and beat more.

### Time inconsistency

The generic meaning of "time inconsistency" is that an agent at one point in time forms or communicates an intention to do something at a later time, yet when that later time arrives has no incentive to carry it out, *with no changes having occurred but the sheer passage of time*. In other words, the intention itself is internally flawed or incoherent. The incoherence reflects an inability to project (Chapter 14).

There are two specific mechanisms that can generate time inconsistency, one intrapersonal and the other interpersonal. The first is illustrated by non-exponential discounting of the future (Chapter 6), and the second by the non-credible threats and promises that I have discussed in this chapter and to which I return in Chapter 24. Although these two sources of time inconsistency have little in common, some of the remedies are the same. Specifically, people can use extrapsychic precommitment devices both to protect themselves against their future selves and to get their way with other people. When people use precommitment devices against themselves, it is sometimes because they anticipate that they might succumb to passion or other visceral factors, and sometimes they believe that they might undergo preference reversal as the result of the sheer passing of time. The failure to stop smoking or drinking illustrates the former case, that of beginning to exercise or to save the latter. When they use precommitment devices in interaction with others, it may be to enhance the credibility of threats and promises, but also for the reasons I spell out in the beginning of the next chapter.

Time inconsistency must be distinguished from time *inconstancy*, and notably from changing time preferences. Formal preferences no less than material preferences are subject to change under the influence of external or internal causes. As one grows older, food preferences may change because of reduced sensitivity to one of the four or five basic tastes. As part of the same

process, time preferences may be affected if people take account of the fact that their life expectancy changes as they grow older. At age x, life expectancy tables tell me that I can expect to live until age x + y and that my chances of living until age x + 2y are very small. Hence I allocate my expected income so as not to have any money left at age x + 2y. However, when I actually reach age x + y, my chances of living until age x + 2y have increased so much that I revise my plans, to exhaust my funds at age x + 3y. This is not inconsistency, but a simple consequence of the fact that I know that I shall die but not when. In other words, changing consistent plans can mimic inconsistent plans. Inconstancy can be debilitating, however, for people whose life is a succession of short-lived long-term plans. Most readers will have come across such persons.

### Bibliographical note

A good elementary introduction to game theory is A. Dixit and S. Skeath, *Games of Strategy*, 2nd edn (New York: Norton, 2004). Among more advanced treatments, I suggest F. Vega-Redondo, *Economics and the Theory of Games* (Cambridge University Press, 2003). An encyclopedic survey with many applications is R. Aumann and S. Hart, *Handbook of Game Theory with Economic Applications*, vols. I–III (Amsterdam: North-Holland, 1992, 1994, 2002). Applications to specific topics are found in J. D. Morrow, *Game Theory for Political Scientists* (Princeton University Press, 1994), and in D. Baird, H. Gertner, and R. Picker, *Game Theory and the Law* (Cambridge, MA: Harvard University Press, 1994). A classic study of conventions is D. Lewis, *Convention* (Cambridge, MA: Harvard University Press, 1969). It is largely inspired by another classic, T. Schelling, *The Strategy of Conflict* (Cambridge, MA: Harvard University Press, 1960), in which the idea of focal points was first expounded. Schelling's work also provided the intuitive foundation for the "second generation" of game theory, formally developed in R. Selten, "Reexamination of the perfectness concept for equilibrium points in extensive games," *International Journal of Game Theory* 4 (1975), 25–55. For various precommitment techniques in political games, see J. Fearon, "Domestic political audiences and the escalation of international disputes," *American Political Science Review* 88 (1994), 577–92. For their use in wage bargaining, see my *The Cement of Society* (Cambridge University Press, 1989).

# 19 Games and behavior

## Intentions and consequences

The conceptual structure of game theory is illuminating. Does it also help us *explain behavior*? Consider the game-theoretic rationale for burning one's bridges or one's ships. This behavior could be undertaken for the strategic reasons set out in the last chapter, but also for others. In the most famous example, Hernán Cortés destroyed all his ships after arriving on the coast of Mexico in 1519, partly to prevent a conspiracy among some of his men to seize a ship and escape to Cuba, partly to add sailors to his infantry. He later wrote that by this act he gave the men the "certainty that they must either win the land or die in the attempt." To my knowledge, there is no evidence that he also intended to signal this fact to Montezuma, as the game-theoretic rationale would require.

In fact, there is *no* documented instance (once again, to my knowledge) of such reasoning.[1] The claim that William the Conqueror burned his ships upon arriving in England in 1066 seems to be a myth. Hume cites James Lancaster's attack in 1594 on Pernambuco in Brazil: "As he approached the shore, he saw it lined with great numbers of the enemy; but no-wise daunted at this appearance, he placed the stoutest of his men in boats, and ordered them to row with such violence, on the landing place, as to split them in pieces. By this bold action, he both deprived his men of all resource but in victory, and terrified the enemy, who fled after a short resistance." The reference to the terror-struck enemy is missing in other accounts, but even if accurate would not support, in fact would contradict, the idea that the enemy fled as a rational response to a rational action. (It is also possible that Lancaster had a prudential fear that the men might retreat out of visceral fear if they were left the opportunity to do so.) Nor do other famous ship-burning episodes fit the game-theoretic pattern. When Agathocles, a tyrant of Sicily in the third century BC, burned his ships, it was because he did not want them to fall into the hands of the enemy, and

[1] Nor does there seem to be any instance of an admiral or general burning his ships or bridges as a precommitment measure to prevent *himself* (rather than his soldiers) from succumbing to fear.

because he could not spare the men to guard them. According to Gibbon, Emperor Julian destroyed the bridges behind him "to convince the troops that they must place their hopes of safety in the success of their arms" and, on another occasion, burned his fleet because it was "the only measure which could save that valuable prize from the hands" of the enemy. Gibbon also cites the example of the Norman adventurer Robert Guiscard, who urged his followers to burn their boats "to deprive cowardice of the means of escape."[2]

I cannot vouch for the accuracy of any of these accounts, which I cite only for methodological purposes. Even though the game-theoretical reasoning might explain the seemingly paradoxical behavior of an agent throwing away some of his assets, *evidence about intentions* is needed to make it more than a just-so story. Game theorists routinely refer to William the Conqueror and to Cortés to illustrate the claim that the enemy *might* rationally refrain from attacking a general who has made himself unable to retreat, but they never (to my knowledge) try to show that anyone ever *did* use that strategy. I return to this general point in the Conclusion.

In some cases, to be sure, there is evidence that the actors *intended* to bring about the consequences predicted by the model. In Chapter 24 I give examples from the theory of bargaining (a branch of game theory), such as firms that build up a large inventory to make it more costly for workers to strike. Proven cases of this kind are, however, surprisingly rare. As I noted in Chapter 3, uncovering the intentions and beliefs of social agents can be difficult, but it is not impossible. If scholars shy away from the difficulty, their work will suffer.

Game theory can also address situations in which some of the actors do not *care* about the consequences. Consider, for instance, the interaction between the European Union and the new entrants from Eastern Europe. The old member states might be tempted to impose conditions for entry that would entail permanently lower agricultural subsidies for the new states, compared to those of same-size old states. In material terms, the new states would benefit so much from entry that they would be better off as second-rank members than as non-members, although less well off than they would be as full members. In psychological terms, the insult of being treated as inferior might cause them to reject such conditions.[3] Anticipating this reaction, the old states might be induced to offer entry on terms of full equality. The belief that material terms were not all the new states cared about might get them better material terms.

[2] In addition to burning ships or bridges, Montaigne offers this example: "There are several examples in Roman history of captains … who would order their horsemen to dismount to remove from the soldiers any hope of flight."

[3] In 2003 President Chirac offered an example of this attitude when he responded to expressions of support for US policy in Iraq by East European politicians by saying that they had missed a great opportunity to keep silent, adding that they had obviously not been very well brought up.

Since I was not privy to the entry negotiations, these remarks are conjectural. We know, however, that arguments of this kind were made at the Federal Convention in Philadelphia in 1787 in the debate over the terms of accession of future western states. Gouverneur Morris and others proposed that these should be admitted as second-rate states, so that they would never be able to outvote the original thirteen states. Against this view, George Mason argued strongly for admission with the same rights as the original states. First, he argued from principle: by admitting the western states on equal terms, the framers would do "what we know to be right in itself." To those who might not accept that argument, he added that the new states would in any case be unlikely to accept a degrading proposal.

> If the Western States are to be admitted into the Union, as they arise, they must be treated as equals, and subjected to no degrading discriminations. They will have the same pride & other passions which we have, and will either not unite with or will speedily revolt from the Union, if they are not in all respects placed on an equal footing with their brethren.

Mason refers to the "pride and passions" of the new states, not to their self-interest. Even if it would in fact be in their interest to accede to the union on unequal terms rather than remain outside, they might still, out of resentment, prefer to stay outside. At the same time, he appeals to the self-interest of the old states, not to their sense of justice. In the terminology of Chapter 4, he is telling them that because the new states might be motivated by *passion rather than by interest*, it would be in the interest of the old states to act as if they were motivated by *reason rather than by interest*.

This situation has been studied experimentally by means of the Ultimatum Game and Dictator Game (Figure 19.1). In the Ultimatum Game, one person (the Proposer) can propose a distribution (x, 10–x) of $10 between him and another person (the Responder). Offers can be made only in whole dollars. If the Responder accepts, that distribution is implemented. If the Responder rejects the proposal, neither gets anything. Although the game has been studied in many variants, I focus on one-shot interactions under conditions of anonymity. Because subjects interact through computer terminals they do not know the identity of their partner. Often it is also made clear to them that the experimenter will be unable to determine who made which choices, thus eliminating the possibility that their decisions might be influenced by the desire to please her. When subjects play the game many times, they never meet the same partner, thus allowing for learning but not for reputation building. Under these conditions, there is maximal scope for the decisions to reflect unfettered self-interest.

Assuming that both agents are rational, are self-interested, and have full information about the pay-off structure and that these facts are common

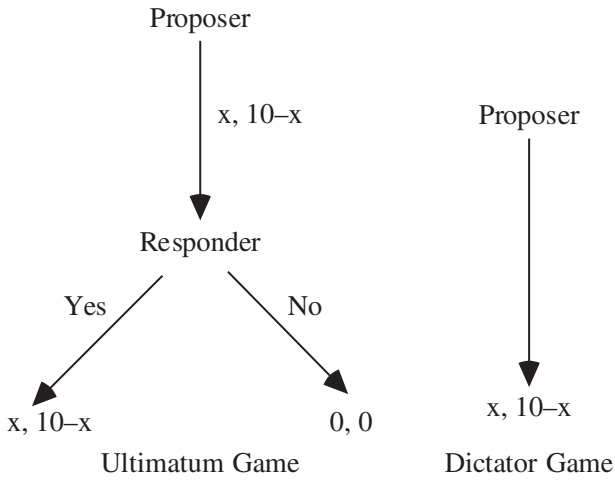



Figure 19.1

knowledge, the Proposer will offer (9, 1), which the Responder will accept. If offers could be made in cents, the Proposer would offer (9.99, 0.01), which would still be accepted, since something is better than nothing. In experiments, proposals are typically around (6, 4). Responders usually reject proposals offering them 2 or less.[4] They are willing to cut off their nose to spite their face. Clearly, one of the assumptions is violated. By virtue of the way the experiment is set up, we can exclude lack of information and lack of common knowledge about information. We cannot exclude, however, failure of rationality or non-self-interested motivations.

The Proposer might be an *altruist*, who prefers a somewhat equal allocation to one in which he gets everything. Although altruism toward perfect strangers who are not in any obvious need is a somewhat strange idea, it is at least consistent with rationality. We can reject this hypothesis, however, by comparing behavior in the Ultimatum Game to behavior in the Dictator Game. In the latter, which is not really a "game" at all, the Proposer unilaterally allocates the money between him and the Responder, leaving the latter no opportunity to respond. If proposals in the Ultimatum Game were dictated only by altruism, allocations in the Dictator Game should be no different. In experiments, though, Proposer behavior is much less generous in the Dictator Game.

[4] I say "typically" and "usually" because of the considerable variation in the findings. Some of the variation is gender-based, some culture-based. Because most experiments use students as subjects, nobody to my knowledge has looked at age differences.

Clearly, the behavior of the Proposer in the Ultimatum Game is driven, at least in part, by the expectation of rejection of ungenerous offers.

To explain this rejection, we might assume that Responders will be motivated by *envy* to reject low offers, and that self-interested Proposers, anticipating this effect, will make offers that are just generous enough to be accepted. If this explanation were correct, we should expect that the frequency of rejection of (8, 2) should be the same when the Proposer is free to propose any allocation and when he is constrained – and known to be constrained – to choose between (8, 2) and (2, 8). In experiments, the rejection rate is lower in the latter case. This result suggests that Responder behavior is determined by considerations of *fairness*. For the Proposer to offer (8, 2) when he could have offered (5, 5) is seen as more unfair than when his only alternative was one that was equally disadvantageous to him. What matter are *intentions*, not outcomes.

This interpretation is supported by the importance of strong reciprocity in other games such as the Trust Game (Chapter 20). People are sometimes willing to punish others, at some cost and no benefit to themselves, for behaving unfairly. This practice seems to violate one of the canons of rationality enumerated in Chapter 14: in a choice between acting and doing nothing, a rational agent will not act if the expected costs exceed the expected benefits. Explanations in terms of altruism or envy would not violate this principle. For an altruist, the outcome can be better when he benefits another at some cost to himself, and for the envious person when he harms another at some cost to himself. Such behavior violates the assumption of self-interest, but not the rationality assumption. By contrast, the fairness explanation seems to violate the latter. Strong reciprocity induces behavior similar to what we do when we stumble over a stone and kick it in retaliation: it does not help and just aggravates the pain.

### Backward induction

In the Ultimatum Game, the game shown in Figure 18.5 and other sequential games, the equilibrium is found by backward induction. In the Ultimatum Game, the Proposer anticipates how the Responder might react to a given proposal and then adjusts his behavior accordingly. In these examples, the calculations involved are very simple. In other experiments, subjects might have to carry out longer chains of reasoning. Two subjects may be told, for instance, to go through three rounds of offers and counteroffers to divide a sum of money, which shrinks 50 percent for each round of offers.[5] At each point, an agent can either accept the proposal and go "right" or make a counterproposal

[5] The shrinking may be seen as an effect of time discounting (Chapter 6).

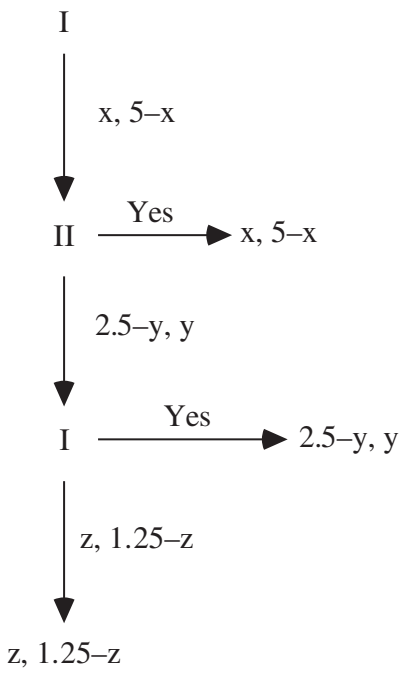



Figure 19.2

and go "down." Rationality, self-interest, and common knowledge then induce the following reasoning.

The person making the first proposal (Player I) will have to take into account whether Player II will prefer the proposed division to one in which he would get a larger share of a smaller pie. At the same time, Player I knows that Player II will not make a proposal that would make I worse off by accepting it than he would be by going to the last round. In Figure 19.2, Player I can get at least 1.25 by taking all that is left in the third round. Player II cannot, therefore, offer him less than 1.25 in the second round, leaving 1.25 as the maximum for himself. Knowing this, Player I will offer (3.75, 1.25) and II will accept.

In experiments, the mean offer made by I is (2.89, 2.11), substantially more generous than the equilibrium offer. Clearly, one or more of the assumptions is violated. (1) The first player might be altruistic. (2) He might fear that the other player would reject the equilibrium offer because he is incapable of following the logic of backward induction. (3) He might himself be unable to follow that logic. (4) He might fear that the second player would reject the equilibrium offer out of resentment. The first, second, and fourth hypotheses can be eliminated by observing the responses when subjects in the role of the first

player are told that they are playing against a computer that is programmed to respond optimally. In that case the average first offer is (3.16, 1.84), which remains substantially more generous than the equilibrium. Since the subjects making the high offers could hardly have altruistic feelings toward a computer or believe it to be incompetent or resentful, they must be incompetent themselves.

It is not that the task is difficult. Once subjects have the logic of backward induction explained to them, they perform impeccably in further games. Rather, the experiment shows that this kind of reasoning does not come naturally to human beings. Even simple forward-looking reasoning may not occur spontaneously, as shown by the Winner's Curse (Chapter 13). The "younger sibling" syndrome (Chapter 17) has some of the same flavor. It is not that people cannot understand, on reflection, that others are as rational and capable of deliberation as they are themselves, only that their spontaneous tendency is to think about others as set in their habits rather than as adjusting to their environments.

### Some failures of rational-choice game theory

Among many other findings that reveal the predictive failures of game theory, I shall discuss the "finitely repeated Prisoner's Dilemma," the "Chain Store Paradox," the "Centipede Game," the "Traveler's Dilemma," and the "Beauty Contest."

When subjects play many successive PDs against one another and know which round will be the last, we observe a substantial proportion of C choices, often exceeding 30 percent. An intuitive explanation is that a player may choose C in one round in the hope that the other will reciprocate ("tit-for-tat"). Yet if the players adopt backward induction, they will understand that in the final game both will choose D since there will not be an opportunity to influence behavior in a later game. In the penultimate game, too, the players will choose D since the behavior in the final game is driven by the previous argument. This argument "zips back" all the way to the first round, thus inducing defection in all games.

A chain store has branches in twenty cities, and in each city it faces a potential challenger. The challenger has to decide whether to set up a store to share the market with the chain store or stay out of the city. The chain store has the option of responding aggressively by predatory underpricing, thus bankrupting the rival but also imposing a loss on itself, or agreeing on market sharing. The pay-offs are as in Figure 19.3; the first number in each pair is the pay-off to the potential entrant.

Backward induction in a single game yields (5, 5) as the equilibrium outcome: the rival enters and the chain store accepts sharing the market. Yet,

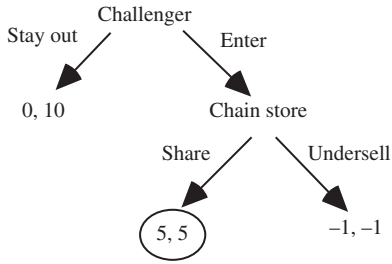



Figure 19.3

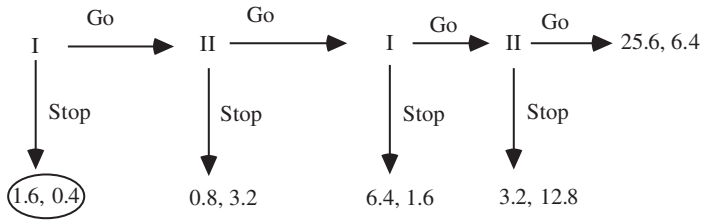



Figure 19.4

thinking ahead to later challenges, the chain store might decide to behave more aggressively and ruin the entrant, at some cost to itself, to deter potential entrants in other cities. But if we apply backward-induction reasoning to the sequence of twenty games, this strategy is not viable. In the twentieth game, there are no further benefits to be had from behaving aggressively, so the firm might as well share the market with the entrant. But that implies that there are no benefits to be gained from predatory pricing in the nineteenth game either, and so on, back to the first game. Although the extent of predatory pricing in actual markets is controversial, it does show up in experimental markets.

The Centipede Game[6] is shown in Figure 19.4 (pay-offs in dollars). Backward induction tells Player I to choose Stop at the beginning, leaving each of the two with one-sixteenth of the pay-offs they could have obtained by continuing all the way to the end. In one typical experiment, 22 percent chose Stop at the first choice "node," 41 percent of those who remained chose Stop at the second node, 74 percent of those still remaining then chose Stop at the third node, and of the remaining, half chose Stop at the fourth node and half chose Go. The deviation from the (circled) equilibrium predicted by backward induction is large, as is the average increase in gains for the players.

[6] The name refers to a version of the game with 100 nodes.

To explain these instances of apparently irrational cooperation and predation, one might stipulate the existence of *uncertainty* about some aspect of the games. Real-life players are rarely faced with a finite and known number of rounds. Often, they might believe that the interaction will continue for an indefinite time, so that there is no final round from which backward induction could begin. In such cases, mutual adoption of tit-for-tat may be an equilibrium in the iterated PD. (It is not unique, since "Always defect, always defect" is also an equilibrium. Structurally, this is a bit like the Assurance Game, with one good equilibrium and one bad.) If real life induces tit-for-tat behavior, agents may apply it to laboratory situations in which it is not optimal.

Alternatively, an agent might be uncertain about the *type* of player she is facing. Suppose there is common knowledge that there are some irrational individuals in the population. It is known that some agents will always cooperate, that others use tit-for-tat in finitely iterated PDs, that still others will use predatory pricing to deter entrants even in the twentieth city, and so on. It is not known, however, exactly who these individuals are. Any agent might, with some positive probability, be irrational. In the Chain Store Paradox, a potential entrant will be deterred if the probability she assigns to the chain store manager's being irrational is sufficiently large. The manager, knowing this, has an incentive to engage in predatory pricing toward the first entrant to make others believe he is irrational. When potential entrants in other cities observe this behavior, they use Bayesian reasoning (Chapter 13) to assign a higher probability to his being irrational. It may not be high enough to deter them, but if he does it again and again, it may eventually reach a level at which it is more rational for them to stay out. A similar argument might explain cooperation in the finitely iterated PD and the Centipede Game.

Yet another possibility is that in the iterated Prisoner's Dilemma and the Centipede Game cooperation has something of a focal-point quality. Although *rational* individuals would defect on the first occasion, *reasonable* persons would not. Although this suggestion (about which more later) is pretty vague, it rings truer, to me at least, than the arguments based on uncertainty about the other person's type. For one thing, these arguments require players to carry out enormously complicated calculations, which take up many pages in textbooks. For another, introspection and casual observation suggest that we do not, when making decisions in everyday life, think about others in this way. When I trust somebody with a small amount of money, but not with a large one, it is not because I assign a small probability to his being unconditionally trustworthy, but because I judge that he can be trusted only when the stakes are not very high.

In the Traveler's Dilemma, two players simultaneously state claims, between \$80 and \$200, for lost luggage. To discourage excessive claims, the airline pays each traveler the minimum of the two claims, adds a sum R to the

person who made the lower claim, and deducts the same amount from the person who made the higher claim. Consider a pair of claims such as (100, 150), yielding pay-offs of (100 + R, 100 – R). This pair cannot be an equilibrium, since the first player would have an incentive to claim 149, yielding a pay-off of 149 + R, to which the second player would respond by claiming 148, and so on. As this example suggests, the unique equilibrium occurs when both claim 80. In experiments, this outcome is in fact observed when R is large. When R is small, however, subjects make claims closer to the upper limit of 200. Again, my intuition is that something like focal-point reasoning is operating. Each traveler knows that given the gains from coordinating on a high claim it would be silly to adopt the equilibrium strategy, and she expects the other to know it too.

John Maynard Keynes compared the stock market to a Beauty Contest. He had in mind contests that were popular in England at the time, in which a newspaper would print 100 photographs, and people would write in to say which six faces they liked most. Everyone who picked the most popular face was automatically entered in a raffle, in which they could win a prize. Keynes wrote, "It is not a case of choosing those [faces] which, to the best of one's judgment, are really the prettiest, nor even those which average opinion genuinely thinks the prettiest. We have reached the third degree where we devote our intelligences to anticipating what average opinion expects the average opinion to be. And there are some, I believe, who practise the fourth, fifth and higher degrees."

In a game inspired by Keynes's remarks, subjects are asked to pick a number between 0 and 100. The player whose number is closest to two-thirds of the average of all the numbers chosen wins a fixed prize. The average is constrained to be 100 or less, implying that two-thirds of the average is constrained to be 67 or less. Hence for any average resulting from the choices of the other players, 67 will be closer to two-thirds of that average than will any number larger than 67. But when numbers are constrained to be 67 or less, two-thirds of the average is constrained to be 44 or less, and so on, until one reaches the unique equilibrium of 0. In experiments, very few subjects choose 0; the average number is around 35. For someone to choose this number, he must believe that most others choose larger numbers – the younger sibling syndrome. The fact that this number is about two-thirds of the average of the whole range, 50, suggests that the typical subject might believe that others pick a number at random while he is free to optimize. Alternatively, the typical subject might believe that others go through two rounds of elimination, leaving him free to optimize by adding a third round.

I have been suggesting that when people fail to conform to the predictions of game theory, it may be because they are *less than rational* or *more than rational*. The younger sibling syndrome is certainly a failure of rationality,

as is the inability to carry out simple backward induction. To be reasonable is to transcend the traps of rationality – to concentrate on the fact that both players can gain while ignoring the best-response logic. As I have said, the latter idea is somewhat akin to the focal-point notion, but only somewhat. Focal points are equilibria, whereas cooperation in the finitely repeated Prisoner's Dilemma, a high claim in the Traveler's Dilemma, or the choice of Go in the Centipede Game are not. What these choices have in common with focal-point choices is a hard-to-define and highly context-dependent property of obviousness and reasonableness.

This argument might seem more similar to magical thinking (Chapter 7) than to focal-point reasoning. To ignore the Sirens of rationality is to follow John Donne's injunction in "The Anniversary":

Who is so safe as we? where none can do

Treason to us, except one of us two.

True and false fears let us refrain.

To ignore *true* fears seems irrational, or magical. (The same holds for ignoring true prospects of gain.) Alternatively, and this is how I prefer to view it, such behavior reflects a higher standard than mere rationality. These are difficult issues, and readers are invited to make up their own minds. Some of the questions are pursued in the next chapter.

### Bibliographical note

C. Camerer, *Behavioral Game Theory* (New York: Russell Sage, 2004), is the source for most of the examples in this chapter. A useful analysis of the conditions under which the predictions of standard game theory break down is J. K. Goeree and C. A. Holt, "Ten little treasures of game theory and ten intuitive contradictions," *American Economic Review* 91 (2001), 1402–22. The apparently simple idea of backward induction turns out to harbor deep paradoxes, some of which are set out in the Introduction to my *The Cement of Society* (Cambridge University Press, 1989). The game illustrated in Figure 19.2 is taken from E. Johnson *et al*., "Detecting failures of backward induction," *Journal of Economic Theory* 104 (2002), 16–47. The Traveler's Dilemma is taken from K. Basu, "The traveler's dilemma: paradoxes of rationality in game theory," *American Economic Review: Papers and Proceedings* 84 (1994), 391–5. For the distinction between the reasonable and the rational, see R. D. Luce and H. Raiffa, *Games and Decisions* (New York: Wiley, 1957), p. 101.

# 20 Trust

### Lowering the guard

Egoism, said Tocqueville, is "the rust of society." Similarly, it is often said that trust is "the lubricant of society."[1] Everyday life would be impossibly difficult if we could not trust others to do what they say they will do, at least to some extent. Although scholars have defined trust in various ways, I shall use a simple behavioral definition: to trust someone is to lower one's guard, to *refrain from taking precautions against an interaction partner*, even when the other, because of opportunism or incompetence, could act in a way that might seem to justify precautions.[2] By "opportunism" I mean shortsighted or "raw" self-interest, unconstrained by either ethical or prudential considerations. Typical opportunistic acts that may justify others' taking precautions include telling a lie, cheating on an exam, shirking at the workplace, breaking a promise, embezzling money, being unfaithful to one's spouse, or choosing the non-cooperative strategy in a Prisoner's Dilemma.

One may or may not trust *oneself* to keep a bargain, stay away from alcohol, or keep the ship on a steady course when the Sirens are calling. Distrust of oneself is revealed by precommitment or by the construction of private rules (Chapter 15). These strategies can be costly, however, because of signaling effects. If others observe one instance of such precautionary behavior toward my future selves, they may infer, incorrectly as we saw in Chapter 12, that I lack self-control in general. Hence they may be reluctant to trust me on occasions when (1) my lack of self-control could be costly for them, (2) no precommitment devices are available, and (3) private rules are irrelevant, as they would be in a one-shot encounter. In many societies, there are norms against total abstention from alcohol as well as norms against drunkenness (Chapter 21).

[1] In an older literature on economic development, *corruption* was sometimes assigned the role of lubricant.

[2] Trust thus understood involves a *double abstention*, one party's refraining from precautions in the hope that the other will refrain from opportunistic behavior.

*Distrust* can take one of two forms. On the one hand, one may simply abstain from interacting with a potential partner when the interaction would make one vulnerable to incompetence or opportunism. On the other hand, one may engage in the interaction but take precautions against these risks. Trust, therefore, is the result of two successive decisions: to engage in the interaction and to abstain from monitoring the interaction partner. Because the decision to abstain from interaction is hard to observe, one may easily underestimate the amount of distrust in society. One might easily think there is more distrust in a society where people are constantly keeping tabs on each other than in one where they largely keep to themselves. Yet on closer inspection one would find that the latter is very inefficient because of the many mutually beneficial bargains that are never struck.

Montaigne described one trusting response: "When I am on my travels, whoever has my purse has full charge of it without supervision." Other instances of showing trust may involve refraining from acts such as the following:

- reading one's spouse's diary
- using proctors to monitor students during exams
- checking the credentials of a prospective employee
- asking for a deposit from a tenant
- insisting on written and legally enforceable contracts
- asking a less wealthy partner to sign a prenuptial agreement
- hiding money from one's children
- locking one's front door when leaving the house
- precommitting oneself to punish defectors in a Prisoner's Dilemma
- asking for a second medical opinion or a quote from a second car mechanic.

As noted, the object of trust can be other people's *ability* or their *motivation*. The distinction is vividly illustrated in the history of resistance movements. In the German-occupied countries during World War II, it happened from time to time that resistance members were killed because they were thought to be German agents. It also happened, although more rarely, that they were killed because they could not be trusted to hold their tongue. A person might turn out to be a drunkard and be executed by the resistance so that he would not reveal dangerous information when drunk. To take a more mundane example, I might question a car mechanic's skill or I might question his honesty. When I ask for a second medical opinion it is often because of worries about the first doctor's competence, although I might also be concerned that she is recommending needless surgery to line her own pocket.

The distinction between trusting someone's ability and trusting her honesty is fundamental, but somewhat neglected in the scholarly literature. The reason is perhaps that competence is more easily observed than honesty, and hence

seems to pose fewer problems. Often, though, one has to be competent to assess competence. Also, a competent person can exert herself to a greater or lesser degree, for opportunistic reasons. Below I mostly discuss trust in honesty.

### Reasons for trust

There are a number of reasons why people may refrain from taking precautions.[3] (1) The cost of taking precautions might exceed the expected benefits, either on a given occasion or over life as a whole. If there is a car mechanic in my village and I would have to travel fifty miles by taxi to get a second quote, it might not be worth it. More generally, life is too short always to fear one might be taken advantage of. The occasional loss that results from trusting the untrustworthy is small compared to the peace of mind that goes with lack of worry. (2) The very act of taking precautions can provide information that can be exploited by opportunists. Montaigne cites the Latin saying *Furem signata sollicitant. Aperta effractarius praeterit* (Locked houses invite the thief: the burglar passes them by when they are wide open). (3) The idea of taking precautions might be incompatible with the agent's emotional attitude toward the other person. When people are in love, they may refuse to engage in the cool calculation involved in a prenuptial agreement. The verse from Donne cited in Chapter 19 is appropriate in this context too: "True and false fears let us refrain." (4) I might have prior beliefs about the trustworthiness of the other person. (5) I might try to *induce* trustworthiness by trusting him or her. (6) Not trusting another person, even if she is a stranger, would show a lack of respect that is inconsistent with a social norm about how to behave toward strangers.

In the following I shall focus on (4) and more briefly on (5) and (6). While many scholars define trust exclusively in terms of (4), I believe the focus on deliberate restraint has the advantage of highlighting the *interaction* between the truster and the trustee. If the trustee *perceives* the lack of precautions, the perception might cause him to act differently than he would otherwise have done. In case (2) this happens because he infers that there will be no *occasion* for opportunistic behavior. In other cases, to be discussed later, the perception may change his *motivation* to behave opportunistically, reflecting a preanalytical intuition that trust has a certain self-fulfilling quality. The same is true of

[3] I use "refrain" to indicate a *deliberate* abstention. In some of the cases I discuss, the idea of taking precautions, for example, by reading the diary of a spouse, may never have crossed the person's mind. Yet this is not "blind trust" as I shall define it later, if the agent had the *opportunity* of taking precautions. For the other person in the relationship, the fact that the agent had the opportunity but did not use it is a telling sign, whether or not the abstention is perceived as deliberate.

distrust. As Proust noted, "As soon as jealousy is discovered, it is considered by its object as a lack of trust which gives her a right to deceive us."

### Reasons for trustworthiness

People may be perceived as *trustworthy* on a number of different grounds. I shall discuss four: past behavior, incentives, signs, and signals. Often, we know – or believe we know (see Chapter 12) – from observation of other people that they consistently keep their promises, abstain from lying, treat property that is not their own carefully, and so on. Moreover, a person who knows himself or herself to be (un)trustworthy will tend to think others are (un)trustworthy too (the so-called false consensus effect) and therefore tend to (dis)trust them. As La Bruyère said, "Knaves easily believe others as bad as themselves; there is no deceiving them, neither do they long deceive."[4] There is experimental evidence that this mechanism does in fact operate. Conversely, A may trust C because he knows that B, whom he trusts, trusts C. The inference may not be valid, however, because B's trust in C might simply be due to the false consensus effect. As these examples show, we often trust or distrust people for bad reasons, believing others to be either more like us or "more like themselves" (more consistent in their behavior) than they in fact are.

In the small international community of diamond merchants, where the temptation for opportunistic behavior is enormous, a verbal agreement without witnesses is as binding as a written contract. A merchant who violated an agreement might pocket a temporary gain, but would be shunned *ever* afterward by *all* other merchants.[5] He would also be unable to pass on the business to his children, as is often the custom in the diamond community. In the case of New York diamond merchants, most of whom live in ultra-Orthodox Jewish communities, a cheater would also suffer social ostracism. The latter mechanism cements the trustworthiness but is not necessary for it. The incentive to maintain a reputation for honesty and trustworthiness is often sufficient.

Signs are *features* of individuals that are thought, rightly or wrongly, to indicate trustworthiness. In a study of what makes taxi drivers willing to trust their passengers not to rob or assault them, women were perceived as more trustworthy than men, older people more than younger, whites more than blacks, the wealthier more than the poorer, the self-absorbed more than the

[4] La Rochefoucauld thought differently: "People are never deceived so easily as when they are out to deceive others."

[5] This is not the simple tit-for-tat mechanism, in which a player who defects in one round and is punished in the next round may be forgiven if he resumes his cooperative behavior. Rather, it is a "grim trigger" mechanism by which a single defection precludes redemption by good behavior later on.

inquisitive, the candid more than the shifty. A Spanish taxi driver in New York would find Spanish passengers more trustworthy than those belonging to other ethnic groups. Catholic drivers in Belfast would find Catholic passengers more trustworthy than Protestants, and vice versa for Protestant drivers. More generic features are having eyes that are not too closely set to each other and looking one's interlocutor in the eyes.

Signals are *behavior* that provides evidence of trustworthiness. These may include the deliberate production or mimicking of signs. For instance, it appears that a good way to generate a frank look is to focus on the root of the nose of one's interlocutor. In this case, the signal will work only if the other person believes that a frank look is a reliable indicator of behavior and ignores how easy it is to fake it. Other behavior works as signals if it is too costly for untrustworthy individuals to afford it. To forge a signature successfully, long practice may be required, whereas writing one's own signature is essentially costless. A poor man might dress up as a Wall Street banker to appear trustworthy to the taxi driver but is unlikely to do so since the costs would be greater than what he could expect to get from the robbery. By contrast, waving the *Wall Street Journal* to hail a taxi is something anyone can afford and hence does not discriminate between the trustworthy and untrustworthy. To the extent that trust relies on the belief that the interaction partner has a long time horizon (a low rate of time discounting), costly displays of physical fitness and slimness can serve as a signal, given the (false) belief that farsightedness as a character trait obtains either across the board or not at all.

Often, we trust people because we perceive them to be motivated not only by their self-interest. Sometimes, however, we trust people only if we see them as self-interested. In *The Maltese Falcon*, Mr. Gutman tells Humphrey Bogart, "I don't trust a man who doesn't look after himself." Napoleon said that Talleyrand was not be trusted because he never asked any favors for his family. As president of France, François Mitterrand was said to be similarly distrustful of those who never asked him for favors. Lyndon B. Johnson was said not to understand or trust men with limited ambitions. More generally, a major problem for confidence tricksters and swindlers is to make their victim believe that they are acting out of self-interest. Suppose I walk up to someone and tell him that there is a fortune to be made by investing a small amount of money up-front. His first question will be "Isn't this too good to be true?" His second question will be "If it really is true, why do you want to share the opportunity with me rather than taking all for yourself?" The successful con artist is able to elicit the trust of his victim by telling a plausible story to explain why he is induced by self-interest to give up part of the gains. In the absence of a prior history of interaction, claims of benevolent motivation are not credible.

Just as people can be (perceived as) more or less trustworthy, they may be more or less *trusting*. That is, if both A and B have the same beliefs about C (or no beliefs at all), A may trust C and B may not. The propensity to trust others is especially important in getting cooperative ventures off the ground. In repeated interactions, cooperation can be sustained by reciprocity *except in the first round*, where there is no prior history of interaction. To get it started, the parties must cooperate unconditionally in the first round. A trusting individual would follow "tit for tat": cooperate in the first round and reciprocate in later rounds. As the proverb has it, "Fool me once, shame on you; fool me twice, shame on me." A distrustful person would follow "tat for tit": defect in the first round and reciprocate in later rounds.

### Experiments on trust

The "Trust Game" (TG) is among the most frequently studied games in behavioral economics, along with the Ultimatum Game (Chapter 19) and public good games (Chapter 23). Typically, one subject (the "investor") has the option of transferring some of her funds (provided by the experimenter) to another subject (the "trustee"). The experimenter then multiplies any amount sent, so that, for instance, if the investor transfers 10 monetary units (MU), the trustee receives 30. The trustee then has the option of transferring some of the gains back to the investor. Rational and selfish trustees would never send anything back; anticipating this fact, rational and selfish investors would not make any transfers. Yet both could be made better off if the investor trusted the trustee, and the latter rewarded the trust. In the example, the trustee could send back 20 MU, leaving both himself and the investor with a net gain of 10. Note that since the experimental set-up usually involved anonymity of investors and trustees to each other, investors have no reason to believe or disbelieve that the trustee is trustworthy, thus excluding explanation (4) of why people trust each other. Trust games take place among strangers.

In one experiment, investors transferred on average around two-thirds of their endowment to the trustee, and trustees made on average a slightly larger back transfer. The larger the forward transfer, the larger the back transfer. These findings are consistent with a number of motivational assumptions, *except* the hypothesis that both agents are motivated by material self-interest and know each other to be so motivated. On that hypothesis the investor, expecting a zero back transfer, would make a zero forward transfer. Typically, however, investors make a positive transfer and trustees a positive back transfer. In some cases, the motivation of the trustees may be altruism or fairness, if they return more than they received. In some cases, however, they return no more and no less than what they received. The motivation behind this

behavior is puzzling. I conjecture that trustees return the same amount they received because they do not want to appear – to themselves or to the investor – as exploitative. As I argue in the next chapter, even the thought of appearing as unfair in the eyes of another person, with whom they will never interact again, is aversive.

Another puzzle concerns the behavior of the investors. In one experiment, subjects were given two risky choices. In one, they were given a chance to gamble $5 to win $10 from the draw of a ball from an urn, told that the urn contained 100 balls, and asked to state the minimum number of winning balls the urn would have to contain to entice them to gamble the $5. The average number was about 64, showing risk aversion (Chapter 13). In the other, they were given the option of investing $5 in a trust game, knowing that the trustee would have the choice between returning $10 and keeping the money to himself. Although only 53 percent thought the trustee would return $10, 71 percent decided to make the investment. This behavior is inconsistent with risk aversion. According to explanation (6), investors make a transfer because not doing so would show disrespect for the trustee, even at "zero acquaintance."

When investors make a transfer, they do not necessarily show trust in my sense of the term. If they do not have the *opportunity of taking* precautions, they cannot *refrain from taking* them. A trust game that focuses on this variable was played in two conditions. In both, the investor could transfer any amount of his 10 MU to a trustee, who could make a back transfer of any amount of the augmented sum (30 MU). In the inappropriately labeled (or so I shall argue) "trust condition," an investor who decides to make a transfer also has to state the amount he wishes the trustee to transfer back to him. In the "incentive condition," the investor is also given the option of stating, at the time he makes the transfer and announces the desired back transfer, that he will impose a fine of 4 MU on the trustee if he transfers back less than the desired amount. Some investors use this option, others do not. If they do not, the trustees know that the investor had the option but refrained from using it. The largest back transfers are made in the incentive condition when trustees are told that no fine will be imposed and the smallest in the same condition when they know that a fine will be imposed, back transfers in the trust condition being at an intermediate level. This effect was anticipated by investors, who invested about 30 percent more in the "incentive, no fine" condition than in either of the others.

The "incentive, no fine" condition corresponds to my definition of trust. What the experimenters call the "trust condition" I would rather call *blind trust*. It is manifested when precautions are *excluded*, as distinct from the case in which they are *not chosen*. The striking finding is that (non-blind) trust induces more cooperation than blind trust. Lowering your guard makes a

difference. In most trust games, though, subjects do not have the opportunity of taking precautions. I conjecture that in most real-world situations of trust they do.[6] If I am right, the relevance of some of the experimental findings may be limited.

I now consider two further experiments that involve the physiological dimensions of trust. The first studied investment size as a function of the presence or absence of the hormone oxytocin. The hormone was known to stimulate pro-social behavior in rodents and to promote the release of breast milk in human females, but the finding that it also promotes pro-social behavior, or trust, in humans came as a surprise. When receiving the hormone, the percentage of investors who transferred their whole endowment to the trustee increased from 21 percent to 45 percent. Three further findings are intriguing. First, trustees who received the hormone did not make larger back transfers. Second, investors given the hormone had the same beliefs about the trustworthiness of trustees (i.e. expectations about back transfers) as those not given the hormone. Third, when investors were told that the back transfers were generated by a random mechanism with the same distribution of pay-offs as when they played against a real person, oxytocin made no difference to the size of transfers. The natural interpretation is that the hormone affected the behavior by making the investors less "betrayal-averse" rather than by making them less risk-averse. The importance of betrayal aversion is also confirmed by other experiments that do not rely on physiological manipulations.

The second experiment, which allowed investors to punish ungenerous trustees, studied what went on in their brains as they were punishing. In this TG, the investor had the choice between transferring his whole endowment of 10 MUs to the trustee and transferring nothing. If he made a transfer, it was quadrupled by the experimenter, leaving the trustee with a total of 50 MU – an original endowment of 10 plus the 40 generated by the investment. The trustee then had the choice between transferring 25 of the 50 back to the investor and transferring nothing. The three possible outcomes, in other words, were (10, 10), (25, 25), and (0, 50) (see Figure 20.1).

In addition, after the trustee made his decision both players received an additional endowment of 20 MU. The investor could use his endowment to punish the trustee, in either of two conditions.[7] In a "costly" condition, the investor could attach up to 20 "punishment points" to the trustee; each point

[6] Referring to the servant who had charge of his purse, Montaigne wrote that "He could cheat me just as well if I kept accounts, and, unless he is a devil, by such reckless trust I oblige him to be honest." Since he asserts that any precautions he could have taken would have been useless, Montaigne did not show trust in the sense I have defined it here, but "blind trust." I suspect, though, that he could have prevented his servant from cheating him by taking very strict precautions.

[7] The trustees were unaware of the punishment option.

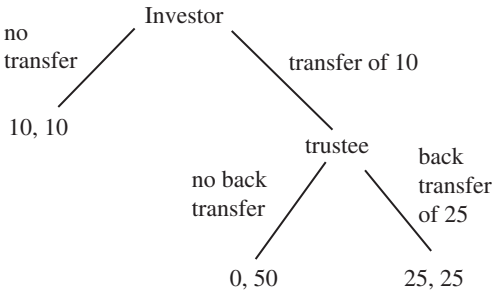



Figure 20.1

caused the investor to lose 1 MU and the trustee to lose 2 MU. Thus by punishing maximally, the investor could ensure that the pay-off of the trustee was reduced from 70 (50 + 20) to 30, while his own was reduced from 20 to 0. In a "costless" condition, only the trustee was affected by the punishment.

All of fifteen investors but one consistently chose to make transfers. The experiment was manipulated so that each investor played against seven trustees, of whom three made the back transfer while four kept all for themselves. These selfish trustees were the focus of the experiment. After a trustee had announced his decision to keep all for himself, the investor had one minute to deliberate and decide whether he wanted to punish the trustee and how severely. During this period his brain was scanned to detect activities in the various regions that might be relevant. One region, the caudate nucleus, is closely linked to the processing of rewards. Another, the prefrontal and orbitofrontal cortex, is linked to the integration of separate cognitive processes, for example, to trade-offs between costs and benefits. In each of these regions the pattern of activities confirmed the hypothesis about the motivation for punishment that I shall now go on to state.

In both the costly and the costless conditions there was a correlation between activation of reward-related circuits and the actual monetary punishment that was imposed. This correlation could mean either that the decision to punish induces satisfaction or that the expected satisfaction from punishment induces the decision to punish. To distinguish between the two hypotheses the experimenters considered eleven subjects who imposed the maximal feasible punishment in the "costless" condition. Among these subjects, those whose reward circuits were more highly activated also imposed more severe punishments in the "costly" condition. As they got more of a kick from punishing, they were willing to spend more on it, thus supporting the second hypothesis. This interpretation is also confirmed by the fact that the cortex was more highly activated in the costly condition, when subjects had to trade off the material costs and psychic benefits of punishment against each other, than in

the costless condition. This finding seems to confirm the "warm glow" theory of this particular form of altruistic behavior (Chapter 5).

I am agnostic about the robustness of the finding. Neuroeconomics still has a long way to go. In its current state, it may be a form of premature reductionism. To be sure, we can say without looking that the brain must be involved, but the exact form of the involvement may elude us. We can compare the case with Descartes's mechanistic physiology, on which Pascal commented in the following terms: "Descartes. We must say summarily: 'This is made by figure and motion,' for it is true. But to say what these are, and to compose the machine, is ridiculous. For it is useless, uncertain, and painful."[8] Also, in real-life situations we tend to avoid rather than punish those who betray our trust (see also Chapter 23).

### *Trust and institutions*

Citizens and institutions may entertain mutual relations of trust and distrust. In many countries, surveys indicate the level of trust citizens have in various institutions. A representative sample is shown in Table 20.1.

This survey, like the others I have seen, does not distinguish between trust or confidence in the *competence* of these institutions and confidence in their *honesty* (see the discussion of virtue and ability in Chapter 5). Because this distinction is fundamental, it is hard to know how to interpret the answers to the survey questions. One might have trust in the competence of the army, but be afraid that it might stage a coup; in fact, the trust in its competence might fuel the fear. Also, survey answers are intrinsically less reliable than behavioral evidence. Individuals show their distrust of banks when they keep their savings at home in the form of cash. Women show their distrust of the police or of the justice system when they fail to report being raped. Some workers refuse to join unions, and some citizens do not watch TV or read newspapers because they distrust either the competence or the honesty of these institutions. People can show their distrust of the political system by not voting, and their distrust of lawyers by trying to reach private settlements. It is harder, however, for individuals to show their distrust of institutions by monitoring them, as distinct from not engaging with them. A fortiori, it is difficult to trust institutions by abstaining from monitoring them.

Relations of trust and distrust between citizens and institutions work in both directions. Public agencies may trust citizens by abstaining from monitoring their claims to welfare services, or distrust them by carrying out invasive home checks. The number of people employed by the Internal Revenue Service as a

[8] This criticism was one of the passages Valéry had in mind when he claimed that Pascal was envious of Descartes (see Chapter 16). Yet since the criticism was obviously right, there is no need to invoke envy!

Table 20.1 *The Harris Poll. February 16–21, 2010. N = 1,010 adults nationwide. Margin of error ± 3.*

"As far as people in charge of running [see below] are concerned, would you say you have a great deal of confidence at all in them?"

| | Percentage |
|---|---|
| **The military** | 59 |
| **Small business** | 50 |
| **Major educational institutions, such as colleges and universities** | 35 |
| **Medicine** | 34 |
| **The US Supreme Court** | 31 |
| **The White House** | 27 |
| **Organized religion** | 26 |
| **The courts and the justice system** | 24 |
| **Public schools** | 22 |
| **Television news** | 17 |
| **Major companies** | 15 |
| **Organized labor** | 14 |
| **The press** | 13 |
| **Law firms** | 13 |
| **Congress** | 8 |
| **Wall Street** | 8 |

percentage of taxpayers may perhaps serve as an indicator of the trust of the government in the honesty of the citizens. The Norwegian practice, established in 2001, of making the income and tax payments of all citizens accessible on the internet seems to reflect a distrust of taxpayers' honesty. The system generated additional tax payments of approximately $100 million (in a country of five million inhabitants), perhaps because citizens fear that neighbors who can observe a discrepancy between reported income and lifestyle will report them to the tax authorities. Hence the distrust may well have been justified. After the system was revised in 2014 so that taxpayers will be automatically informed of the identity of those who have accessed their tax information, some of these additional payments may be lost.

An important open question is whether increased government trust in citizens will lead to increased citizens' trust in the government. If the police, for instance, behave in ways that makes them appear less of an enemy to citizens, will the latter be more willing to volunteer information? If the claims of women to have been raped are treated with less skepticism, will more claims be made?

### Bibliographical note

Evidence that the trustworthy are those who trust others is offered in D. Glaeser *et al.*, "Measuring trust," *Quarterly Journal of Economics* 115

(2000), 811–46. The diamond merchant community in New York is analyzed by B. Richman, "How community institutions create economic advantage: Jewish diamond merchants in New York," *Law and Social Inquiry* 31 (2006), 382–420. The use of signs and signals by taxi drivers to determine the trustworthiness of their passengers is the topic of D. Gambetta and H. Hamill, *Streetwise* (New York: Russell Sage, 2005). The ways to make a scam appear credible are analyzed in a neglected book by N. Leff, *Swindling and Selling* (New York: Free Press, 1976). On trust games in general, see C. Camerer, *Behavioral Game Theory* (New York: Russell Sage, 2004), Chapter 2.7. The trust game with optional punishment is reported in E. Fehr and B. Rockenbach, "Detrimental effects of sanctions on human altruism," *Nature* 422 (2003), 137–40. The social-norm explanation of why investors make a transfer even when they have no reason to expect a back transfer is proposed in D. Dunning *et al*., "Trust at zero acquaintance," *Journal of Personality and Social Psychology*, 107 (2014), 122–41. The impact of oxytocin on trust is shown in M. Kosfeld *et al*., "Oxytocin increases trust in humans," *Nature* 435 (2005), 673–6. The idea of betrayal-aversion is confirmed by I. Bohnet and R. Zeckhauser, "Trust, risk and betrayal," *Journal of Economic Behavior and Organization* 55 (2004), 467–84. The study of trust and revenge is by J. F. de Quervain *et al*. (2004), "The neural basis of altruistic punishment," *Science* 305 (2004), 1254–8.

# 21 Social norms

### The collective consciousness

Sociologists sometimes refer to the "collective consciousness" of a community, the set of values and beliefs shared (and known or believed to be shared) by its members. On the value side, the collective consciousness includes moral and social norms, religion, and political ideologies. On the belief side, it includes opinions about factual matters as well as about causal relations, ranging from rumors about the white slave trade to beliefs about the perverse effects of unemployment benefits. In this chapter I consider social norms and their operation. In the next chapter, I consider modes of collective or, better, interactive belief formation. There is a double asymmetry in my treatment of values and beliefs. On the one hand, I have little to say about the emergence of social norms, not because the question is uninteresting but because I find it too hard. On the other hand, I have little to say about the substance of popular or collective beliefs. Their content varies greatly in time and space, whereas the mechanisms of emergence, propagation, change, and collapse of beliefs are more invariant.

### The operation of social norms

Consider two statements:

> Always wear black clothes in strong sunshine.
> Always wear black clothes at a funeral.

The first injunction is a matter of instrumental rationality, since the air between the body and the clothes circulates more rapidly when the garments are black. The second expresses a social norm, which has no obvious instrumental significance. The existence and importance of social norms cannot be doubted. The proximate causes involved in their operation are reasonably well understood. Yet their ultimate origin and function (if any) remain controversial.

A social norm is an injunction to act or to abstain from acting. Some norms are unconditional: "Do X; do not do Y."[1] They include the norms not to eat human flesh, not to have sexual intercourse with a sibling, not to break into the queue, never to wear red clothes (as some mothers tell their daughters), to wear black clothes at a funeral, to begin with the outermost knife and fork and work inward toward the plate, to treat the sickest patient first (even if the chances of curing him are worse than those for other patients). Other norms are conditional: "If you do X, then do Y," or "If others do X, then do X." In many groups, there is a norm that the person who first suggests that some action be taken is then charged with carrying it out;[2] as a result, many good suggestions are never made. A childless couple may feel subject to a norm that whoever first suggests they have a child will have a larger share in raising it; as a result some couples who would like to have a child may remain childless.[3] There may not be a norm telling me to send Christmas cards to my cousins, but once I begin there is a norm to continue and another norm telling my cousins to reciprocate. Yet although conditional, these norms are not conditional on any *outcome* to be realized by the action, as is the injunction to wear black in strong sunshine.

Revolutionary action by crowds can also be subject to social norms: it is legitimate to destroy, but not to steal; to kill, but not to rape. Tocqueville commented on the efficacy of such norms in the revolutions of 1789 and 1848. Concerning the first, he wrote that "if among the armed men someone committed a base action, he was immediately jailed by his comrades. This is a peculiar feature of our French people." Concerning the 1848 Revolution, of which he had been an eyewitness, he wrote that the ban on theft did "not prevent a lot of robbery on such days, for ... there are always rascals everywhere who jeer at the morality of the main body and are very contemptuous of its conception of honor *when nobody is looking*." In other words, those motivated by moral norms can act as enforcers of a social norm with the same content.

More examples will be offered later. First, however, I need to say something about what lends causal efficacy to social norms and how they differ from other norms. A simple response to the first question is that social norms operate through informal *sanctions* directed at norm violators. Typically, sanctions affect the material situation of the offender, either by the mechanism of direct punishment or by the loss of opportunities caused by social ostracism.

[1] In the following, "conditional" and "unconditional" refer to the content of the social norm. As noted in Chapter 5, all social norms are conditional in the sense that their operation is contingent on the presence of an observer.

[2] This norm may be linked to focal-point reasoning.

[3] Hence the norm induces a game of Chicken.

A farmer who violates community norms may see his barn burned down and his sheep disemboweled. Alternatively, he may find his neighbor denying his request for help with the harvest. The mechanism of *gossip* can act as a multiplier on these sanctions, by adding third-party sanctions to the original second-party punishment.

Consider what a cattle farmer might do when a neighbor's cattle repeatedly trespass on her land. She may seize the cattle, at a benefit to herself and at some cost to the neighbor. She may destroy the cattle or reduce their value (e.g. by castrating a bull), at no benefit to herself and some cost to the neighbor. She may herd the offending livestock to some distant place, at some cost to herself and to the neighbor. Or she might cut off all relations with the neighbor (ostracism). The last response could be inefficient, however, in that it might not deter future trespasses. The first response might be seen as an aggressive taking rather than as a punishment. The second and especially the third responses are more adequate, in that they clearly indicate an intention to punish, if need be at some cost to the punisher.

In general, however, I believe that ostracism or avoidance is the most important reaction to norm violations. If instead of repeated trespass the neighbor had engaged in a one-shot break of promise, cutting off relations would have been the more natural reaction. This claim is supported by the general idea that social norms operate through the *emotions* of shame in the norm violator and of contempt in the observer of the violation (Chapter 8). Because the action tendency of contempt is avoidance, which often causes material losses for the ostracized person, there is a link between the emotional response and the imposition of sanctions. Yet the sanctions are often more important as a vehicle for the communication of emotion than they are in their own right. Moreover, the cost of sanctioning *to the sanctioner* may be especially important in communicating the strength of his emotion.

The sanction theory of social norms runs into an obvious problem: what motivates the sanctioners to punish? What is in it for them? Typically, sanctioning is costly or risky for the sanctioner. Even if he does not give up an opportunity for mutually profitable interaction, the expression of disapproval might trigger an angry and even violent reaction in the target. There is an important distinction here between spontaneous disapproval and deliberate shaming. The latter can easily backfire, causing the target to be angry rather than ashamed. Even when the disapproval is in fact spontaneous, the target may, perhaps self-servingly, interpret it as intentional shaming and react accordingly. For this reason, sanctioning is a risky business. Why, then, do people engage in it?

One answer might be that non-punishers themselves risk punishment. This no doubt happens. In a society with strong norms of revenge one might expect that a person who fails to shun someone who fails to take revenge would

herself be shunned. Among schoolchildren, a child might be more willing to interact with a "nerd" when not observed by classmates. Yet a child who abstains from joining the mob in harassing a child who is friendly to the nerd is unlikely himself to be harassed. Hence the third-party harassers are not likely to be motivated by the fear of punishment. Experimentally, the question might be examined by seeing whether third parties would punish Responders who, by accepting very low offers in the Ultimatum Game, fail to punish ungenerous Proposers. I would be surprised if they did, and even more surprised if fourth-party observers punished non-punishing third parties. At a few removes from the original violation, this mechanism ceases to be plausible.[4]

A more parsimonious and adequate explanation of sanctioning relies on the spontaneous triggering of contempt and the associated action tendency. Anger, too, may be involved, because of the fluid distinction between social and moral norms. Also, *flaunting* one's violation of social norms is likely to trigger anger rather than contempt because it tells other people that one does not care about their reactions. Although these spontaneous action tendencies may be kept in check by the costs and risks of sanctioning, they may be capable of overriding the latter. Ostracizing the nerd who could help his classmates with homework is costly, as was the refusal of aristocrats under the *ancien régime* to let their daughters marry wealthy commoners. When a "taste" for discrimination takes the form of refusing to employ or buy from members of despised minority groups or women, economic efficiency may suffer. Often, such behavior reflects the operation of social norms rather than of idiosyncratic individual preferences, as shown by phrases such as "Jew-lover" or "nigger-lover" used to condemn those who go against the norm.

Sanctioning requires *information* about the norm violation, either by direct observation or by hearsay, notably gossip. In large communities such information may be hard to come by. For this reason, it may be hard to enforce the social norm of voting in national elections. To overcome this difficulty, one might publish the names of non-voters. In a large-scale field experiment, a substantially higher turnout was observed among those who received mailings promising to publicize their turnout to their household or their neighbors. In this study, the citizens automatically received information about who voted and who did not. Alternatively, one might leave it to the citizens whether to seek out this information, by posting the names of non-voters on the internet. This practice already exists in Argentina, where it is combined with a fine for non-voting. The "naming, shaming and blaming" produced by publicity could also be a *substitute* for fines, and perhaps a more effective method.

[4] In the Old South, things were organized more rigorously. "Peer pressure demanded that every young man assist the [slave] patrols and take turns in whipping the suspects."

### What social norms are not

Social norms need to be distinguished from a number of related phenomena: moral norms, quasi-moral norms, legal norms, and conventions. Although the dividing lines may be fluid, there are clear-cut cases in each category. Both moral and quasi-moral norms (Chapter 5) are capable of shaping behavior even when the agent believes herself to be unobserved by others. By contrast, the shame that sustains social norms is triggered by the perceived contempt of others. The corresponding action tendency is to escape from their accusing stares: to hide, run, and even kill oneself.

Legal norms differ from social norms in that they are enforced by specialized agents who typically impose direct punishment rather than ostracism, experiments with legal "shaming" notwithstanding. Legal and social norms interact in numerous ways. In 1990, for instance, some state legislators in Louisiana pushed for reduction of criminal sanctions applicable to informal punishers of flag burners. Even after an edict of 1701 allowed the French nobility to engage in commerce (only wholesale, not retail), it was more than fifty years before they overcame the social norms prohibiting the practice. In some communities, there are social norms against appealing to legal norms, whereas in others people litigate at the drop of a hat.

Conventions, or convention equilibria, can in principle be enforced through the sheer self-interest of the agent, without any action by others. As noted in Chapter 18, they are often quite arbitrary. At the first day of a conference, each participant may find his or her seat more or less randomly. On the second day, a convention has been created: people converge to their chosen seats because doing so is the obvious (focal-point) allocative mechanism. On the third day, the convention has hardened into an entitlement: I get angry if another participant has taken "my" seat. Yet although the social norm cements the arbitrary convention and makes it more likely to be respected, it is not indispensable. Among New Yorkers, there is a convention to celebrate New Year's Eve in Times Square, but since few people would know whether a given person showed up or not, there is little opportunity for sanctioning. Even if the norm of driving on the right side of the road were not reinforced by social norms and legal norms, the dangers to the driver of switching into the left lane would be a strong deterrent.

A complex category is that of unwritten legal and political norms such as constitutional conventions.[5] These are usually not legally enforceable, although courts may take account of them in decisions. Instead, they are

[5] The phrase "constitutional convention" is used about two entirely different aspects of constitutions: the unwritten norms that supplement the written constitution and the constituent assemblies that are often used to adopt a written constitution.

enforced by political sanctions, or the fear of such sanctions. Until 1940, for instance, the American constitutional convention that nobody could serve as president more than twice was enforced by the belief that anyone who tried to do so would be defeated. This was why Ulysses Grant did not stand for a third term. Such norms, of which there are many, have some of the flavor of social norms, since they are enforced by the diffuse force of public opinion rather than by specialized agencies.[6] Other political conventions are better seen as equilibria in repeated games. In many parliamentary systems there is, for instance, a convention that when an administration leaves office its internal documents are sealed and become available (to historians) only after several decades. Although any given administration might be tempted to open the archives of its predecessors and use them as political ammunition, the knowledge that this would set a precedent for its successor to do the same is sufficient to deter it from doing so. This is not a convention in the sense of Chapter 18, since each administration would prefer to deviate from it as long as others do not.

### Norms and externalities

There are norms against those who impose small negative externalities on many others (Chapter 17). When people litter in the park, spit in the street, urinate in the lake, or drink from the office coffee pot without dropping a quarter into the cup, they usually try to do so unseen. Even when they do not actually fear sanctions, the mere thought that others might think badly of them may deter them from performing these actions when observed. Norms of this kind are socially useful in the strong sense that they make *everybody* better off. The norm against spitting in public places is an especially good example. Before one knew how contagious diseases were spread, spitting was a perfectly acceptable practice and widely catered to by spittoons. Once the mechanism of contagion was understood, "No spitting" signs appeared in many public places. Today, the norm is so entrenched (in some countries at least) that the signs have been taken down.

[6] Two American examples illustrate the force of these norms. When T. Roosevelt stood for a third term after a split in the Republican Party, which had failed to nominate him, feelings ran high. While on one of his speech-making tours, he was shot at by a man of unbalanced mind, who said: "I shot Theodore Roosevelt because he was a menace to the country. He should not have a third term. I shot him as a warning that men must not try to have more than two terms as President." Regarding the norm (which remains unwritten in about half of the American states) that all members of the state electoral college must vote for the candidate who received the greatest number of votes, one observer predicted that "an Elector who failed to vote for the nominee of his party would be the object of execration, and in times of very high excitement might be the subject of a lynching."

In this example, we can observe the norm emerging and claim with some confidence that it came about *because* it was in the public interest. The danger was perceived, a legal norm was created, and the social norm followed. Whether the perception of negative externalities can create social norms *without* the intermediary step of public intervention is more questionable. The mere fact that a norm is needed, and perceived to be needed, does not automatically bring it about. In developing countries, there is no social norm to limit family size. Social norms against overgrazing and overfishing have not emerged spontaneously to prevent the tragedy of the commons. There is no norm regulating the use of antibiotics, although their excessive use imposes externalities on others through the development of more resistant micro-organisms. Norms against playing music on the public beach and against using cell phones in the concert hall also owe (I conjecture) their origin to action by the relevant authorities. Over and over again, we find that outside intervention is necessary to stop people from imposing these negative externalities on each other. In some cases, as in the norm against spitting, people may refrain even when the legal norm disappears or ceases to be enforced. In others, such as China's "one-child" policy, it seems unlikely (but perhaps not impossible) that the behavior would persist if the regulation were to be lifted.

Smaller groups may be able to impose these norms without external intervention. In the workplace, there is often a strong norm against rate busters, because it is believed that their efforts might cause the management to lower the piece rate. (In this case the externality takes the form of an increase in the probability of a rate cut.) Although the management might want to commit itself to a policy of fixed piece rates, to induce workers to make a stronger effort, it may not be able to make a credible promise to this effect. Strikebreakers, too, are often heavily sanctioned by their fellow workers. It is perhaps significant that these two cases involve common opposition to an adversary. In a "game against nature" such as overgrazing, solidarity does not seem to emerge as easily, because free riding is not seen as *betrayal*. In firms that pay workers using individual piece rates, the norm against rate busting may emerge because it is seen as benefiting the "enemy." It would be less likely to arise (except if fueled by envy) in a workers' cooperative.

Other social norms target negative externalities that one group of people imposes on another. The norm against smoking, even in places where it is still legally allowed, is an example.[7] In many Western societies today, guests who smoke often abstain without even asking the host whether they would be allowed to. What one might call "noise externalities" underlie the norm

[7] The most important externality is caused by smoke inhalation (passive smoking). It is sometimes also claimed that smokers impose a negative externality on other smokers who want to quit but cannot resist the desire to smoke triggered by the visual cue of others smoking.

"Children should be seen but not heard." There are two ways in which this injunction could be a social norm and not merely a form of parental punishment. First, children might ostracize other children who violate the norm. Second, parents might ostracize other parents whose children violate it. In train compartments, those who want to impose a "fresh air externality" on others usually lose the contest with those who impose a "stuffy air externality." (On buses in Paris, this norm is posted as an obligation.) The reason may be that closed windows are perceived as the default option and hence as a normative baseline.

### Norms and conformism

Some social norms are little but injunctions *not to stick one's neck out*. Inhabitants of small towns everywhere will recognize the "Law of Jante" written down (in 1933) by one who got away:

> Thou shalt not believe thou *art* something.
> Thou shalt not believe thou art as good as *we*.
> Thou shalt not believe thou art more wise than *we*.
> Thou shalt not fancy thyself better than *we*.
> Thou shalt not believe thou knowest more than *we*.
> Thou shalt not believe thou art greater than *we*.
> Thou shalt not believe *thou* amountest to anything.
> Thou shalt not laugh at *us*.
> Thou shalt not believe that anyone is concerned with *thee*.
> Thou shalt not believe thou canst teach *us* anything.

These norms can have very bad social consequences. They can discourage the gifted from using their talents and may lead to their being branded as witches if nevertheless they use them. Luck, too, is frowned upon. Among the Bemba of Northern Rhodesia, it is said that to find a beehive with honey in the woods is good luck; to find two beehives is very good luck; to find three is witchcraft.

### Codes of honor

Strong and often subtle norms can regulate behavior in feuds, vendettas, duels, and revenge more generally. The norms define the actions that call for a retaliation or a challenge, the conditions under which and the means by which it can or must be carried out, and the fate of someone who fails to live up to the primary norm. Beginning with the last, the failure to take revenge often causes a kind of civic death, in which the agent is completely cut off from normal social relations. Within his family, his opinion counts for nothing; if he

ventures outside his home, he is met with ridicule or worse. It is a paradigmatic situation of contempt, inducing intolerable shame.

Anything that can be seen, however remotely, as an insult to the agent's honor can trigger retaliation. In prerevolutionary Paris, the Vicomte de Ségur, a prominent rake about town, amused himself by writing small epigrams in verse. A rival who was jealous of his reputation wrote a little verse himself subtly mocking Ségur's verses. As revenge, Ségur seduced the rival's mistress and then, when she announced that she was pregnant, told her that he had just been using her to get back at his rival and that now that he had attained his aim he was no longer interested in her. (She subsequently died in childbirth.) He went back to Paris and told the story to anyone who would listen, never encountering disapproval. *Les liaisons dangereuses*, it seems, was but a feeble imitation of reality.

In nineteenth-century Corsica, there were four circumstances that justified or required vengeance: when a woman had been dishonored, when an engagement had been broken, when a close relative had been killed, and when false testimony in court led to the conviction of a member of one's family. In one case, a notary was convicted of homicide on false testimony and subsequently died in prison. His brother became a bandit and over a period of years killed all fourteen prosecution witnesses. These are all cases of vengeance for the sake of *maintaining* one's honor. The system of honor also included, however, actions undertaken for the purpose of *gaining* honor. Montaigne refers to "what is said by the Italians when they wish to reprove that rash bravery found in younger men by calling them *bisognosi d'honore*, 'needy of honour.'"

In the American South people react more strongly to perceived insults than do northerners. Homicide rates are higher in the South, and people express stronger approval of violent reactions to affronts. In an ingenious study, a confederate of the experimenter bumped into the subject, "accidentally on purpose," and called him an "asshole." Afterward cortisol levels (reflecting reactions to the incident) and testosterone levels (reflecting preparation for future aggression) rose dramatically more in southern than in northern subjects. In another experiment, subjects continued walking down the hallway where they had been "bumped" and saw a large football player type (a confederate) walking toward them in a determined manner. The hallway had been cluttered with tables so that there was room only for one person to pass at a time, essentially creating a game of Chicken. Southerners went much closer to the other person (three feet) before they "chickened out" than did northerners (nine feet).

Do codes of honor serve any social function? If they do, can the function explain why they exist? The idea that the practice of revenge is a useful form of population control is too arbitrary to be taken seriously. An alternative view, that norms of revenge provide a functional equivalent of organized law

enforcement in societies with a weak state, is also implausible. The Mediterranean and Middle Eastern societies that have subscribed to these norms have had levels of violence and mortality rates among young men far above what are found elsewhere.[8] As suggested by the observation by Montaigne just quoted, norms of revenge and the larger code of honor in which they are embedded may *light as many fires as they put out*. Often, feuds create more disruption than they control.

Others have argued that norms of honor evolve in sparsely settled herding societies, in which a reputation for willingly using violence serves as a useful, even indispensable, deterrent to theft. The culture of honor in the American South has been explained in this perspective. Over and above the general problems of functional explanation, this analysis runs into the difficulty that codes of honor were equally strong in the court of the French kings in the seventeenth and eighteenth centuries, to name only one non-rural example. Some of those who focus on codes of honor in the urban aristocracy rather than among rural herders then come up with another functional explanation: in the absence of war the nobility "needed" duels to keep up their warlike spirit. If one does not provide a mechanism by which the need would generate its own satisfaction, this argument is worthless. These polemical comments do not imply that I have a better explanation to offer.

### Norms of etiquette

A further set of social norms are those involved in rules of manners or *etiquette*. Codes of dress, language, table behavior, and the like are often relentless in their detail, condemning to ostracism those who miss the smallest nuance.[9] In all societies there is a norm regulating the appropriate distance from other people to maintain on social occasions. If one moves inside the private space of a person (in the United States perhaps fifteen inches) one risks being shunned as uncouth. The norm is unusual, however, in that the individuals concerned are often unaware of its existence and operation. Most norms of etiquette are highly codified, often literally so. They are not only (for the most part) pointless but also sometimes even cruel in their consequences, as when a five-year-old girl goes home in tears because her friends ridiculed her new stroller for her baby doll on the grounds that it *had no brakes*. In

[8] One might object that the relevant comparison is with the level of violence that would obtain in the "state of nature." If that state is defined by an exclusive concern with self-interest and the absence of any state-like agencies, it would not produce violence motivated by envy, spite, or anger. The anticipation of forceful appropriation by the strong of the goods produced by the weak might have a chilling effect on production, thus preventing actual violence from occurring.

[9] In aristocratic societies, *gross* deviations are sometimes accepted, when seen as deliberate rather than as evidence of rule ignorance. Proust's Charlus is an example.

prerevolutionary Paris, a young officer, wealthy but not noble, tried to gate-crash a ball at Versailles. "He was treated so severely that in his despair over the ridicule with which he was covered, at a time when ridicule was the worst of all evils, he killed himself when he came back to Paris."

The puzzle is why these intrinsically trivial matters take on such importance. The disproportionate disapproval triggered by a breach of etiquette may be due to the unfounded belief that people are all of a piece (Chapter 12), so that someone who violates an unimportant norm is likely also to violate more consequential ones. Also, the violation of trivial norms of etiquette may be seen as a non-trivial show of disregard for what other people think. This leaves unexplained, however, why the unimportant norms exist at all.

Functional explanations are very common. The subtle rules of etiquette among the elite exist, allegedly, in order to make it more difficult for outsiders to "crash the party" by imitating the rule-governed behavior. There is no doubt that these rules often have the *effect* of keeping upstarts down, but that does not offer an explanation of why they exist. As many self-proletarianized students have discovered, it is very difficult to break into the working class for someone who was not born in it. In Norway in the 1970s, for instance, young Maoists found that making fun of the royal family was a sure way of alienating themselves from the class they were trying to join. Yet nobody has suggested that the norms of the working class exist *in order to* make it more difficult for outsiders to pass themselves off as workers. The argument makes no more sense for the norms of the elite.

Norms of etiquette can impose heavy costs, while benefiting nobody. In eighteenth-century Massachusetts, social norms required widows and widowers to provide mourners with rings, gloves, and scarves, all of which had to come from England. In 1741, the Massachusetts House of Representatives tried to put an end to these wasteful practices, by forbidding the distribution of scarves and rings and limiting the number of those who could receive gloves to six persons, in addition to the minister and six pallbearers. The legislation was largely ignored. When the conflict with England caused economic depression, it was welcomed by some individuals because, in the words of a contemporary, "each individual being ransomed from the tyranny of fashion, will be free to act as his circumstances may require, and such freedom can scarce be purchas'd too *dear*." In contemporary China, too, many poor families are ruined by norms of highly public gift-giving at funerals and weddings. Even if they get something back when they are hosting an event, ceremonies on average cost more than twice the income from gifts received. There is evidence that the resulting "income squeeze" leads to in utero malnutrition and subsequently stunted growth.

When expensive private schools impose the use of uniforms on the students, it is probably also to reduce the cost of following social norms of dress – and to

remove the visible contrast between those who can afford it and those who cannot. A more interested motive lay behind the sumptuary laws that have been enacted at many times and places, to prevent commoners from imitating the manners of the aristocracy. Here, too, however, the sheer expense mattered. In the Italian town of Lucca, the social norm requiring expensive dowries threatened to reduce the population, because fathers were unable to equip their daughters to make them eligible. The municipality enacted a limitation on the number of expensive dresses per person, but it was not enforceable. These examples serve to undermine a rosy view of social norms as providing a solution to problems that institutions cannot address. The opposite can be true: the norms can be harmful, and institutions incapable of restraining them.

### Norms of drinking

If social norms were invariably geared to enhance the welfare of the individual or of society we might expect them to be directed against heavy drinking that is perceived to have harmful short-term or long-term consequences. There are indeed many norms of this kind. Some norms, usually linked to religion, demand total abstinence. Islam and some Protestant sects have absolute bans on alcohol. Secular norms, by contrast, often enjoin drinking in moderation. The Italian norm "Never drink between meals" has the dual effect of limiting total consumption and of reducing the rate of absorption of alcohol, thus buffering the short-term effect on the body. In Iceland, there are norms against drinking in the presence of the children and against drinking on fishing trips.

Alcohol-related norms do not, however, always enhance welfare. There are norms that condemn abstinence, as well as norms that enjoin people to drink heavily. Among the Mapuche Indians of Chile, drinking alone is criticized, and so is abstinence; such behavior is seen as showing lack of trust. The traditional French culture condemns both the teetotaler and the drunkard. In Italy, distrust of abstainers is expressed in a proverb, "May God protect me from those who do not drink." In youth subcultures of many countries, abstainers are subject to heavy pressure and ridicule. Conversely, there are many societies in which heavy drinking is socially prescribed. In Mexico and Nigeria, the macho qualities shown in the ability to drink heavily are much admired. In pre-revolutionary Russia, excessive drinking was obligatory in the subculture of young officers.

When abstinence is condemned or when heavy drinking is socially mandatory, would-be abstainers may have to resort to subterfuge. In Sweden, a common question is "Do you want sherry, or are you driving?" It is so accepted that abstaining alcoholics often say they are driving because this relieves them of the social pressure that otherwise would certainly be exerted by the host to convince the guest to have a drink. The norm of drinking can

only be offset by another norm (against drunk driving). Similarly, it has been argued that conversions to Protestantism provide an alternative for some Latin Americans who want to opt out of the system of community governance in which even the rituals often involve heavy drinking and drunkenness. Again, the norm of drinking can only be overridden by another norm, which in this case has the backing of religion.

These are cases of the strategic use of norms. Conversely, people can behave strategically to get around the norms. Some ancient Chinese considered alcohol itself to be sacred and drank it only in sacrificial ceremonies; eventually, they would sacrifice whenever they wanted to drink. In Spain, at certain hours, not to drink on an empty stomach is a tacit cultural proscription, so food will be included with the drinking. In both cases, we observe a reverse of the original causal link: rather than obeying the conditional norm of drinking only when they are doing X, people do X whenever they want to have a drink.

### Norms of queuing

The queue is a norm-ridden social system. At the same time, the queue is a transient phenomenon, unlike the other contexts I have discussed. Before I proceed to discuss norms of queuing, let me pursue this contrast for a moment.

It is probably a common intuition that norms have less impact on behavior in communities with high turnover. According to Tocqueville, "Men who live in democracies are too mobile to allow some group of them to establish and enforce a code of etiquette. Each individual therefore behaves more or less as he pleases, and manners are always to some extent incoherent because they are shaped by each individual's feelings and ideas rather than conforming to an ideal model held up in advance for everyone to imitate." Earlier, I referred to the small-town norm of "Don't stick your neck out," with the implication that in more anonymous interactions deviant behavior would be less severely sanctioned. In light of this intuition, it is interesting that as communities grow larger and more mobile, we observe the emergence of *norms regulating the behavior among strangers*. This remark strengthens the interpretation of norms in terms of emotion rather than material sanctions. Under most circumstances, it is difficult to impose tangible sanctions on a person who violates a queue norm. There is not even much space for avoidance. People can give full rein, however, to expressions of contempt or indignation.

There is a norm, I believe, against walking up to the person at the head of a bus queue and offering him or her money in exchange for the place. This norm is obviously inefficient: if the person who is asked accepts and moves to the back of the line in exchange for the money, both persons benefit and nobody is hurt. According to Tocqueville, such norms against open display of wealth in

public are specific to democratic societies: "Do you see this opulent citizen? . . . His dress is simple, his demeanor modest. Within the four walls of his home, luxury is adored." There may also be an underlying idea that the use of queuing is a valuable counterweight to the pervasive use of money in allocating scarce goods. To prevent the rich from getting everything, let some goods be allocated by a mechanism that puts them at a disadvantage, because of their greater opportunity costs of queuing. In Communist Poland, where queuing was endemic, there was no norm against purchasing a place in a queue, probably because this practice was seen as one of many necessary forms of jockeying for position. Other forms included hiring people to stand in line or moving back and forth between several queues while asking people in each of them to "hold the place." There were norms regulating these activities, and deviations were sanctioned. A surprising norm was the rule against reading while queuing. According to an observer, "women do not want to be told, even by implication, that they are actually wasting time in queues. If one reads or works in the queue, this implicitly reminds others that they are wasting time. The response is to scold the deviant, putting the reminder out of sight and mind." In addition, people reading or working would shirk their duty of monitoring violations of queue norms.

A different kind of violation occurs when someone intrudes in a queue, whether at the head of the line or somewhere in the middle. In this case, the negative reactions of people behind the intruder in the queue might be due to considerations of cost, be it in the form of time costs or (if they are queuing for a scarce good) material costs. Alternatively, they might be due to outrage or indignation. Experiments find that although both factors may be at work, subjects usually have a stronger reaction to illegitimate intrusions than to legitimate ones that impose equal costs. (A telling fact in these studies is that the confederates of the experimenter who were asked to intrude in the queue felt the task to be highly aversive.) There is often a norm to the effect that responsibility for rejecting intruders lies with the person immediately behind him or her. There are also norms regulating place holding in queues. In my supermarket, the norm seems to be that it is acceptable to leave the shopping cart in the line to go to pick up one item from the shelves, but not to go back several times. In Australian football queues, the norm in leaving position markers is that one must not be absent for periods longer than two to three hours. Although it might seem more efficient if most people placed a marker in the queue and then went home for a while, this practice would violate equality since the people who remained in the queue to maintain it would be disfavored.

The basic principle of fairness in queuing, "First in, first out," can be violated when there are multiple and independent queues. Thus reported customer satisfaction is higher in the single-queue Wendy's restaurants than in the multi-queue Burger King and McDonald's restaurants, although the

latter average half the waiting time of Wendy's. At Houston airport, customers with checked luggage complained about the baggage delay (a one-minute walk to the carousel and a seven-minute wait at the carousel), compared to passengers with hand luggage who could proceed directly to the taxi stand. When the airport authorities changed the disembarking location so that all customers had to walk for six minutes, complaints dropped to nearly zero.

### The norm of tipping

Tipping for service is not a negligible phenomenon. Estimates of tips in US restaurants range from $5 billion to $27 billion a year; adding tips to taxi drivers, hairdressers, and others would yield a larger figure. Estimates of the fraction of income that waiters derive from tips range from 8 percent (the Internal Revenue Service assumption) to 58 percent for waiters serving full-course meals. In some contexts tipping may seem puzzling, in others less so. If you go to the same hairdresser each time you need a haircut, you tip to ensure good service; the same applies for meals in your favorite restaurant. Tipping in one-shot encounters, such as a taxi ride or a meal in a restaurant you do not expect to visit again, is more puzzling. Such behavior is in fact doubly puzzling: it cannot be sustained by two-party interaction over time, nor by third-party sanctions at the time of the encounter. If you are the only passenger in the taxi, other people are rarely in a position to know whether you tip the taxi driver adequately; nor are other customers in the restaurant likely to notice how much you tip your waiter.

Tipping, it has been argued, is an efficient way of remunerating waiters. It is obviously easier for the client to monitor the quality of service than it is for the restaurant owner. Hence decentralizing the monitoring function and linking reward to observed performance are a way of overcoming the "principal-agent problem" (how to prevent workers from shirking) that besets many contractual relationships (Chapter 25). Tipping, therefore, might be part of an "implicit contract" for the purpose of enhancing efficiency. But as Sam Goldwyn said, an unwritten contract isn't worth the paper it's written on. The argument, like many other attempts to explain social norms, is merely a piece of unsupported functionalism. The idea that restaurant owners who forbid tipping are eliminated in the competition with those who allow it is entirely conjectural, and in any case would not explain why customers do tip in the latter. Also, when assessed empirically, tipping does not seem to pass the appropriate efficiency tests. It does not, for instance, appear to be more prevalent in occupations where monitoring is easier. The fact that waiters often pool their tips also undermines the efficiency argument. Indirectly, however, the pooling of tips could enhance efficiency, if waiters ostracize colleagues who give such bad service that they bring little back to the pool.

I do not know why there is a norm to tip in certain occupations and not in others, or why the same service receives a tip in some countries and not in others. Once a norm exists, however, we can understand why people tip: they simply do not like the idea that others, such as a disappointed taxi driver, might disapprove of them, even if they do not expect to meet them again. Being the object of the contemptuous stare of the other is not necessary. It may be enough simply to know or have reason to believe that the other feels contempt. To take another example, the belief that others might disapprove explains why I abstain from picking my nose on the subway platform when a train is passing by without stopping, even if there are no other people on either platform.

### Why norms?

The importance of social norms for the regulation of behavior and the proximate mechanism by which they operate are, as I said, fairly well understood. I do not believe, however, that we have a good understanding of their origin. There are two separate questions. First, what is the evolutionary origin of the correlative emotions of shame and contempt that sustain social norms? In other words, why are there social norms at all? Second, why do specific norms exist in specific societies? How and when do they arise; how and when do they disappear?

A simple answer to the first question is that we care intensely about what other people think about us. We seek their approval and fear their disapproval. This answer, however, only raises the same question at one remove: why should we care about what other people think about us? In some cases, to be sure, a reputation can be useful and worth cultivating. Yet the thought that the taxi driver might think badly about us if we do not leave a tip is entirely divorced from reputational concerns. Also, since the reason others think badly about us is that we have violated a social norm, explaining norms by the desire that others not think badly about us is to some extent circular.

Concerning the second question, the most common answer is that norms emerge to regulate externalities. There is something to this idea if we add, as I argued we should, that social norms against imposing negative externalities on others are usually ushered in by an outside authority. There is a general social norm to obey the law. If fines were seen as prices, and prison as no more stigmatizing than a stay in a hospital, there would be no such norm, but in general these reactions to lawbreaking are not seen as equivalent to other, objectively equal burdens. People feel ashamed of going to jail and try to hide the fact if they can.[10] When the law bans behavior that imposes negative

[10] In Norway, there used to be a mandatory three-weeks' prison sentence for drunk driving. Some people allegedly took a sunlamp with them into the prison cell, to acquire a tan they could use to buttress their story of having taken a holiday.

externalities on others, the social norm of obeying the law may spill over into a norm against that behavior. The norm may persist even if the law that gave rise to it falls into disuse. This outcome may be hard to distinguish, however, from the emergence of the "good equilibrium" in an Assurance Game (Chapter 18). If the state induces cooperation by punishing defectors and then dismantles the punishment apparatus, people may continue to cooperate because each person's top-ranked situation is the one in which she and everybody else cooperate (there is no free-rider temptation).

With regard to many of the other norms I have discussed, such as the norm against offering money to buy someone's place in the bus queue, norms of etiquette, or norms of tipping, it is harder to come up with an explanation of their emergence and persistence. One line of argument, often offered by economists, is that the persistence of norms can be explained as equilibrium behavior and that their emergence is a matter of accident and history about which social science has little to say. Since an implicit premise of this book is that the dividing line between social science and history is artificial and pointless, I cannot agree with the latter claim. As to the former, I have argued that social norms typically do not exhibit the best-response logic that characterizes strategic games. When, unobserved, I observe another violating a norm, sanctioning the violator is typically not a best response.

### Bibliographical note

This chapter builds on, and (I hope) improves on, the account of norms I proposed in *The Cement of Society* (Cambridge University Press, 1989) and, more succinctly, in "Social norms and economic theory," *Journal of Economic Perspectives* 3 (1989), 99–117. Influential discussions of social norms are J. Coleman, *Foundations of Social Theory* (Cambridge, MA: Harvard University Press, 1990), R. Ellickson, *Order Without Law* (Cambridge MA: Harvard University Press, 199 ), and E. Posner, *Law and Social Norms* (Cambridge, MA: Harvard University Press, 2000). I learned from all of them but was not persuaded by any. For an instructive criticism of Posner, see the review by R. McAdams, *Yale Law Journal* 110 (2001), 625–90. Useful discussions of unwritten constitutional norms or conventions are found in two articles by J. Jaconelli, "The nature of constitutional convention," *Legal Studies* 24 (1999), 24–46, and "Do constitutional conventions bind?" *Cambridge Law Journal* 64 (2005), 149–76. The two American examples are taken from H. Horwill, *The Usages of the American Constitution* (Oxford University Press, 1925). The Law of Jante is taken from A. Sandemose, *A Fugitive Crosses his Trail* (New York: Knopf, 1936). The role of witchcraft in sustaining norms against sticking one's neck out is discussed in K. Thomas, *Religion and the Decline of Magic* (Harmondsworth: Penguin, 1973). I discuss codes of

honor and revenge in Chapter 3 of *Alchemies of the Mind* (Cambridge University Press, 1999). The story about the Vicomte de Ségur is taken from *Les mémoires de la Comtesse de Boigne* (Paris: Mercure de France, 1999), vol. I, pp. 73–4. These memoirs (*ibid*., p. 38) are also the source of the story about the young officer who killed himself out of shame for being ridiculed. Norms of etiquette are the topic of P. Bourdieu, *Distinction* (Cambridge MA: Harvard University Press, 1987), with a distinctly functionalist slant. The experimental studies on "the culture of honor" are reported in R. Nisbett and D. Cohen, *The Culture of Honor* (Boulder, CO: Westview Press, 1996). The examples of norms of drinking are taken from my *Strong Feelings* (Cambridge, MA: MIT Press, 1999). The misadventures of self-proletarianized students in Norway are charted in a wonderfully amusing novel, unfortunately not translated into English, by D. Solstad, *Gymnaslaerer Pedersens beretning om den store politiske vekkelsen som har hjemsøkt vårt land* (Oslo: Gyldendal, 1982). A movie version with English subtitles is available. The story about funeral norms in Massachusetts is told in T. H. Breen, *The Marketplace of Revolution* (Oxford University Press, 2005), and the story about dowries in Lucca in H. Freudenberger, "Fashion, sumptuary laws, and business," *Business History* 37 (1963), 433–49. The effects of costly funerals and weddings in China are documented in X. Chen and X. Chang, "Costly posturing: relative status, ceremonies, and early child development in China," working paper (Yale School of Public Health, 2012). For norms of queuing in Communist Poland, see J. Hrada, "Consumer shortages in Poland," *Sociological Quarterly* 26 (1985), 387–404. For some of the other examples, see L. Mann, "Queue culture: the waiting line as a social system', *American Journal of Sociology* 75 (1969), 340–54. The efficiency-based explanation of the norm of tipping is offered by N. Jacob and A. Page, "Production, information costs and economic organization: the buyer monitoring case," *American Economic Review* 70 (1980), 476–8. It is criticized in M. Conlin, M. Lynn, and T. O'Donoghue, "The norm of restaurant tipping," *Journal of Economic Behavior & Organization* 5 (2003), 297–321, which proposes an account closer to the one sketched here.

# 22 Collective belief formation

### Tocqueville on conformism

The mechanisms of belief formation I considered in Chapter 7 operate for the main part at the level of the individual, in the sense that the beliefs held by one person owe little to those held or expressed by others. In this chapter I discuss some mechanisms of collective or interactive belief formation. To illustrate the distinction, consider Tocqueville's analyses of American conformism. One explanation why Americans tend to have the same ideas is simply that they live under similar conditions: Since "men equal in condition ... see things from the same angle, their minds are naturally inclined towards analogous ideas, and while each of them may diverge from his contemporaries and form beliefs of his own, all end up unwittingly and unintentionally sharing a certain number of opinions in common." Another explanation relies on the pressure to conform: "In America the majority erects a formidable barrier around thought. Within the limits thus laid down, the writer is free, but woe unto him who dares to venture beyond those limits. Not that he need fear an *auto-da-fé*, but he must face all sorts of unpleasantness and daily persecution."

This last passage suggests that people conform outwardly, because of social pressure, but not necessarily inwardly. As he also writes, if you hold a deviant view, "your fellow creatures will shun you as one who is impure. And even those who believe in your innocence will abandon you, lest they, too, be shunned in return." Other passages suggest that conformism reaches all the way to the soul, so that people eventually develop a sincere belief in the majority view. Two mechanisms are suggested, one "cold" or cognitive and another "hot" or motivational. On the one hand, "it seems unlikely ... that everyone being equally enlightened, truth should not lie with the greater number." On the other hand, the fact that "American political laws are such that the majority is sovereign ... greatly increases its inherent influence over the intellect, for there is no more inveterate habit of man than to recognize superior wisdom in his oppressor."

### Experimental findings

I have cited Tocqueville at some length (and shall cite him again in this chapter), because of his acute insights into these matters. The questions he identified – outward versus inward conformism, and cognitive versus motivational mechanisms – are very much with us today. To address them I shall first cite some classic experiments on conformity.

In the most famous experiment, subjects were asked to indicate which of three lines A, B, and C was closer in length to a given line D. There were three conditions: private, doubly public, and singly public. In the private condition, subjects stated their answer when no one else was present, besides the experimenter. In this case 99 percent indicated that D was closest to B, suggesting the unambiguous correctness of this answer. Yet in the two public conditions a substantial minority of subjects gave different replies. In both conditions, the subject answered after several others (confederates of the experimenter) had unanimously said that A was closer in length. In the doubly public condition, in which the subject gave his answer in the presence of the confederates, about one-third agreed that A was closer.[1] In the singly public condition, in which subjects stated their opinion privately after they had heard what the others said, conformism was reduced but not eliminated.

The excess conformism in the doubly public condition was arguably due to *fear of disapproval*. The residual conformism in the singly public condition could be due to *learning* ("so many others are not likely to be wrong") or to *dissonance reduction*. The latter explanation seems the more plausible. Those who conformed privately with the majority are unlikely to have done so on the basis of rational learning only, given the poor cognitive status of the majority view. Some motivational factor must have been at work.

Another experiment strengthens that interpretation. Here the subjects had a more ambiguous task, detecting the distance a light source in a dark room had traveled. Although the source was in fact immobile, isolated subjects judged it to have traveled about four inches (the "autokinetic effect"). Having heard one confederate say that the light had moved between fifteen and sixteen inches, the subjects estimated the distance to be about eight inches. With two confederates making estimates in the sixteen-inch range, the estimate of the subjects was about fourteen inches. The presence of one

[1] In Chapter 5 I discussed how observing others can trigger behavior similar to theirs by the quasi-moral norm of fairness, whereas being observed *by* others may trigger similar behavior through the fear of disapproval. In belief formation, observation by others can also produce conformity through fear of disapproval, whereas conformity produced by the observation of others occurs either by learning or by dissonance reduction.

confederate, that is, led to a four-inch increase in the estimate, and that of a second to a further six-inch increase.

In a process of Bayesian learning (Chapter 13) I can rely on other observers to correct my perception or memory. Their estimates of some fact, such as the distance traveled by the light, can serve to modify my initial assessment. How *much* they will affect it depends on my beliefs about the reliability of their perception and on the number of these other observers. In this experiment, the subject would presumably attach the same reliability to each confederate. Whatever that reliability might be, the change in his estimate caused by one confederate's stating the distance as sixteen inches should be greater than the additional change caused by the second confederate.[2] This contradicts the findings, however, since the second confederate caused a *greater* adjustment than did the first. There seems to be a dissonance-reduction effect, caused by the discomfort of finding oneself disagreeing with the majority, which cannot be reduced to rational learning.

The second experiment had a further, interesting feature. It ran for several "generations," over which the confederates were gradually replaced by naive subjects. Thus in the second generation of a two-confederate experiment, one confederate was replaced by a naive subject from a first-generation experiment, while in the third generation the other confederate was also replaced by a naive subject from an earlier generation. In subsequent generations all participants were naive subjects who had been previously exposed either to confederates or to other subjects who had been exposed to confederates, and so on. The experiment was designed so that the newly inducted subject in each generation spoke after the two others. The designers of the experiment had anticipated that the artificially high estimates would be maintained indefinitely, but were proved wrong. After about six generations in three-person groups and eight generations in four-person groups the estimates converged to four inches, that is, the distance estimate given by isolated subjects. The belief in the emperor's new clothes did not perpetuate itself indefinitely. If some cultural beliefs with poor support in reality do maintain themselves over time, it could be because the discrepancy is hard to observe or because they are supported on other grounds. The use of lotteries to identify good hunting or fishing sites, as is the practice in some societies, may have survived because of their religious significance.

[2] In the numerical example used to illustrate Bayesian learning in Chapter 13, each new piece of confirming evidence brings about a smaller increase in probability than did the previous one. This "decreasing marginal value of new information" is a quite general phenomenon.

### Pluralistic ignorance

At the outset of this chapter I distinguished between two reasons why people at a given time might hold or profess similar beliefs: because they are influenced by similar conditions (correlation) or because they influence each other (causation). A special case of the first is provided by the many examples of simultaneous discoveries, such as the invention of the calculus by Newton and Leibniz more or less at the same time. Although no one knows exactly what the "similar conditions" were in that case, the idea may have been "in the air." For another case of the simultaneous appearance of similar ideas consider the idea of the emperor's new clothes. Hans Christian Andersen's tale was published in 1837. In the second volume of *Democracy in America*, published in 1840, Tocqueville came up with a similar idea to explain the apparent stability of majority opinion:

> Time, events, or individual efforts by solitary minds can in some cases ultimately undermine or gradually destroy a belief without giving any external sign that this is happening. No one combats the doomed belief openly. No forces gather to make war on it. Its proponents quietly abandon it one by one, until only a minority still clings to it. In this situation, its reign persists. Since its enemies continue to hold their peace or to communicate their thoughts only in secret, it is a long time before they can be sure that a great revolution has taken place, and, being in doubt, they make no move. They watch and keep silent. The majority no longer believes, but it still appears to believe, and this hollow ghost of public opinion is enough to chill the blood of would-be innovators and reduce them to respectful silence.

A passage from Tocqueville's *Old Regime* (1856) makes a similar point about religion. In the course of the French Revolution "those who retained their old faith became afraid of being alone in their allegiance, and, dreading isolation more than heresy, joined the crowd without sharing its beliefs. So what was still only the opinion of a part of the nation came to be regarded as the opinion of all, and from then on seemed irresistible even to those who had given it this false appearance." In *Democracy in America*, he made the same argument with regard to the false appearance of *faith*: "The unbeliever ceases to believe in true religion but continues to deem it useful. Looking at religious beliefs from a human angle, he recognizes their power over mores, their influence on laws ... Having lost the faith ... he is afraid to take it from anyone who possesses it still. Meanwhile, the man who continues to believe does not hesitate to expose his faith to the view of all ... With those who do not believe hiding their incredulity and those who do believe showing their faith, public opinion develops in favor of religion."

In these passages, Tocqueville refers to beliefs that people *profess* to hold (or abstain from disavowing), not to beliefs they actually and sincerely hold. In this respect his analysis differs from behavior in the moving-light experiment

and in the singly public condition of the line-matching experiment.[3] This is not, however, a hard and fast distinction. As I have argued in several places, it is not always clear what it means to "believe" that something is the case. Even in the singly public condition, the "belief" of the subjects who said that A was the matching line may have been somewhat faint. They might not, for instance, have been willing to bet money on the proposition. Also, stating a belief may, under some circumstances, induce a tendency to endorse it (Chapter 7).

Modern psychology rediscovered Tocqueville's insight under the heading of "pluralistic ignorance." In extreme cases, nobody believes in the truth of a certain proposition but everybody believes that everybody else believes it. One could imagine, for example, a society in which everybody holds Assurance Game preferences, but believes that everybody else holds Prisoner's Dilemma preferences (Chapter 18).[4] Nobody would cooperate, and each would take the non-cooperation of others as confirmation of their beliefs. Society would be stuck in a bad equilibrium. In more realistic cases, most people do not believe the proposition but believe that most people do. Thus in Israel almost 60 percent agree to Palestinian autonomy in the territories, but only 30 percent realize that this is the majority position. This state of affairs would obviously tend to discourage advocates for Palestinian autonomy, and to maintain a status quo that the majority would like to change. According to joint Israeli–Palestinian opinion polls, a majority of both populations believes that a return to the 1967 borders is the preferred solution. I conjecture that the majorities falsely believe that only a minority on the other side holds this view. If this is so, that discrepancy would also have a chilling effect on the prospects for a durable peace.

These cases differ from the pathological situations in which everybody publicly professes a certain belief while knowing that nobody actually holds it in private. Communism displayed this *culture of hypocrisy* to an extreme

[3] Actually the moving-light experiment was doubly public, leaving scope for insincerity. Yet the ambiguous nature of the task presumably facilitated sincere, or quasi-sincere, adoption of the exaggerated belief. I assume that the reason for the procedure was that the study of successive generations would have been hard to do in a singly public condition.

[4] Tocqueville asserts that relations among occupational groups in the French *ancien régime* had this structure: "Our forefathers lacked the word *individualism*, which we have forged for our own use, because in their day there was no individual who did not belong to a group and who could regard himself as absolutely isolated. But each of the myriad small groups that made up French society had no thought for any group other than itself. There was, if I may put it this way, a sort of collective individualism, which prepared people for the true individualism that we have come to know ... Had anyone been able to plumb their inner depths, moreover, he would have discovered that these same people, being quite similar to one another, regarded the flimsy barriers that divided them as contrary to both the public interest and common sense and, in theory at any rate, already worshiped unity. Each clung to his own condition because others distinguished themselves by theirs, yet all were prepared to merge into a single mass provided no one stood out in any way or was elevated above the common level."

degree, at least in its final gerontocratic stage. Pluralistic ignorance and cultures of hypocrisy can be sustained by the same mechanism, namely, fear of disapproval or punishment for stating deviant views. The difference is that in pluralistic ignorance, the disapproval is horizontal – meted out by fellow citizens who falsely believe they have to ostracize deviants lest they themselves be ostracized. By contrast, the culture of hypocrisy works by vertically imposed punishment: those who do not express enthusiasm for fulfilling the plan or hatred of the class enemy are likely to lose their jobs or worse. The vertical punishment may then induce horizontal measures, if people avoid or punish deviants lest they be punished as deviants themselves. Thus in a description of the Great Soviet Purge of 1937, we read that "Boris Pasternak was in trouble for not signing a collective request from well-known writers for the execution of Kamenev and Zinoviev; the prose writer Iurii Olesha was in trouble for defending Pasternak."

Pluralistic ignorance also differs from the mechanism underlying the passive-bystander syndrome observed in the "Kitty Genovese" killing. In that case, each individual believed that the passivity of others justified his or her own. The cause cannot have been social pressure or a desire to conform to group norms, since the thirty-eight bystanders were too isolated from each other to form a community. Rather, the passivity seemed justified by an *inference:* since nobody else seemed to be doing anything, the situation could not be very serious. The "raw data" (her cries) were overwhelmed by this inference. We shall look more closely at this mechanism shortly. Here I only want to note that the situation did not involve pluralistic ignorance, since there was no discrepancy between what each person privately believed and the beliefs he or she imputed to others.

The culture of drinking has been shown to illustrate pluralistic ignorance. On many American campuses, there is a culture of heavy drinking among undergraduates, especially male. Most students do not feel comfortable with the heavy levels of drinking but go along because they believe, wrongly, that most others do.[5] Their drinking behavior conforms to what they wrongly believe to be the typical attitude on campus rather than to their private attitudes. Another example can be taken from an experiment in which students were told to read an article written in a deliberately obtuse style that made it virtually incomprehensible, and then asked how well they had understood it and how well they thought others had understood it. In one condition, the students had the option of seeking out the experimenter and asking for assistance; in another they were expressly told they could not do so. Even in the former condition, no students went to see the experimenter because the

[5] It may be true, however, both that most students do not drink and that most friends of most students drink, if those who drink have more friends than those who do not.

procedure for doing so required that they risk embarrassing themselves. Each student seems to have believed, however, that whereas he or she stayed put out of fear of embarrassment, others did so because they understood the article and needed no help. Hence students in that condition tended to believe that others had understood the article better than they had themselves. The difference disappeared in the other condition. Conjecturally, this effect might be due to an "older sibling syndrome." As noted in Chapter 17, we are all aware of our own inner anguishes and fears, but since we do not have direct access to the inner life of others, we tend to see them as more mature and self-possessed.

In the study of drinking on campus, it was also found that over time private attitudes, beliefs about the attitudes of others, and behavior moved into line with one another, raising the question of the *stability* of pluralistic ignorance. There are in fact two ways in which it might disappear: by the false beliefs about others becoming true or by people ceasing to hold them. If each person adopts the belief he or she (falsely) imputed to others, that imputation would in fact become true. This would most likely happen by dissonance reduction, caused either by the discomfort of disagreeing with the majority or the discomfort of saying one thing and believing another. This seems to be what happened with drinking on campus.

On the other hand, the situation might unravel.[6] Suppose that 20 percent of group members show in their behavior that they do not hold the belief in question, and that the remaining 80 percent pay lip service to it because they require more than 20 percent of non-conformists in the group to become non-conformists themselves. Specifically, suppose that in a group of a hundred, there are twenty non-conformists, ten who would be willing to "come out" if at least twenty-five have already done so, fifteen who would do so if at least thirty-five have, and fifty-five who would join if at least fifty have shown their true colors. As stated, the majority culture is stable. Imagine, however, that five of the most conformist individuals leave or die and are replaced by five non-conformists. In that case, the majority would unravel. The twenty-five non-conformists would create the conditions for ten more to join them; the resulting thirty-five would attract fifteen more, thus generating the requisite threshold for the remaining fifty to join. Instead of referring to the process as the *unraveling of conformism*, we may also see it as the *snowballing of non-conformism*. We shall observe a similar dynamic in collective action (Chapter 23).

Conformism may unravel in many other ways. The little child in Andersen's tale is reflected in the line-matching experiments: when a *single* confederate

[6] The extinction of false beliefs in the moving-light experiment is also a form of unraveling, due to the fact that in each generation subjects use their own "raw data" to adjust the estimated distance somewhat downward compared to what they hear from others.

stated the veridical opinion that D was the closest in length to B, the conformism all but disappeared. For another example, consider the widespread belief in both England and France prior to the Reformation that the king could heal scrofula by touching the sick person. The Reformation undermined this belief, since Catholics in France and Anglicans in England now were compelled to explain why the evidence in the other country was spurious. But recognizing the possibility of large-scale collective error turned out to be dangerous, since the allegedly invalid proofs used to support the belief in the other country were not very different from the ones invoked in one's own.

Another mechanism for unraveling is the publication of an opinion survey. Prior to the 1972 referendum on Norwegian entry into the Common Market (as it was called then), the government, the main opposition parties, and the major newspapers were all massively in favor of entry. Although, as the referendum showed, there was a popular majority against entry, each individual opponent would have been led to believe himself or herself a member of a small minority had not the opinion polls indicated otherwise. Without the polls, the outcome of the referendum would in all likelihood have been different. Some of those opposed to entry would have abstained from voting, since the outcome would have been seen as a foregone conclusion. Also, the movement that was formed to persuade the undecided would have remained small and uninfluential. In the period between the introduction of universal suffrage and the rise of opinion surveys, the scope for pluralistic ignorance about political matters must have been considerable.

Elections, too, may reveal a state of pluralistic ignorance. Before the first semi-free elections that were agreed upon in the Round Table Talks between the Communist Party and the opposition in Poland in 1989, most people on both sides of the negotiations as well as foreign observers believed that the Communists would obtain enough votes to allow them to stay in power. Although a majority of the population was opposed to the regime, they may have believed that they formed a minority. In fact, it is unlikely that the Communists would have agreed to the institutional reforms had they not been confident of sufficient support.[7] On June 4, the opposition swept the elections, the regime collapsed, and other countries in the region followed suit. I return to some aspect of this snowballing process, within and across countries, in Chapter 23.

[7] I conjecture that the Ministry of the Interior conducted opinion polls that confirmed this belief. Yet in a semi-totalitarian society citizens may be reluctant to express their true opinions, even if assured of anonymity.

## Rumors, fears, and hopes

Another tale by Hans Christian Andersen, "There Is No Doubt About It," illustrates how "one little feather may easily grow into five hens" through successive exaggerations. The study of rumor formation and propagation is not, to my knowledge, very far advanced. With some exceptions, psychologists have not made much progress on the issue, partly because laboratory studies cannot create the tense and dense atmosphere in which rumors are born and spread. To illustrate, let me quote a letter from 1798 about the rumors of a conspiracy in England to assist the French: "I am quite disgusted and discouraged at the difficulty of knowing what is the truth of facts that are even passing almost under one's own eyes. I seldom venture to report serious reports at more than second hand at the farthest – but now I find that even too far & that little less is safe than the testimony of your own eyes – nor do I know that even they ought to be trusted in the first moments of fear acting upon prejudice." The last four words offer a remarkable insight, but hardly one that could be confirmed in the laboratory.

Also, in my opinion, psychologists rely too heavily on speculative ideas about the function of rumors. While rumors may indeed serve the need of finding meaning and order in the universe (Chapter 9), other alleged functions seem more doubtful. Economists tend to view rumors as "rational herding" or "informational cascades." While this approach may sometimes be appropriate, there is little doubt that most rumors have an irrational component. Hence I shall rely mostly on the work by historians, notably on the pathbreaking study of the "Great Fear" of 1789 by Georges Lefebvre.

I shall not try to define rumors, except to note that they are not necessarily false, but, when true, true only by accident. Because of their causal origin in hopes and fears, they cannot be characterized as "justified true beliefs." I shall classify them into two categories, optimistic and pessimistic, related to, respectively, wishful and counterwishful thinking.

### Optimistic rumors

- rumors in the French countryside in the spring and summer of 1789 that the convocation of the Estates-General implied that Louis XVI had already decided to satisfy the demands of the people;
- rumors among Virginian slaves in 1829 that the state constitutional convention had their liberation as its main goal;
- rumors among Galician peasants in 1845 "that their final emancipation was impending, and that, indeed, a decree freeing them had been signed by the emperor but filched by the gentry";[8]

[8] By contrast to these three cases of unfounded optimism, the "If only the King knew" syndrome, implying a conflict between the good ruler and his evil advisers, did not exist under Stalin, at

- rumors among Napoleon's followers about his return after each of his two defeats in 1814 and 1815;
- rumors after the French revolutions of 1830 and 1848 about impending tax reductions;
- rumors among the opponents of Napoleon III about his impending fall after an assassination attempt in 1858;
- rumors that tens of thousands of Russian soldiers had joined the Allied troops in August 1914;
- rumors in America after Pearl Harbor that the Japanese had oil and food reserves for six months only, and that revolutions in Japan and Germany were imminent;
- rumors about rising stock values ("irrational exuberance").

**Pessimistic rumors**

- rumors in France under the *ancien régime* about famines created by speculators;
- rumors in France under the *ancien régime* about the creation of a tax on children;
- rumors in the French countryside in the spring and summer of 1789 about roving "brigands" who were out to cut the grain before it was ripe ("the Great Fear");
- rumors during the War of 1812 that a slave revolt had erupted in the District of Columbia and Maryland;
- rumors among Napoleon's opponents about his return after his defeat in 1815;
- rumors about a socialist leveling in the wake of the 1848 Revolution;
- rumors about a massive invasion of Germany in March 1848 by impoverished French workers, plundering, burning, and killing;
- rumors among the partisans of Napoleon III about his impending fall after an assassination attempt in 1858;
- panics in financial markets;
- rumors in India in 1935 about an impending earthquake;
- German rumors about *francs-tireurs* shooting at German soldiers from the rooftops when Germany invaded Belgium in 1914;
- Belgian rumors about German atrocities when Germany invaded Belgium in 1914;

least not in the 1930s. An historian of the period writes that "No rumors are known to have circulated in the villages concerning Stalin's benevolent intentions or his role in rescuing mice from the predatory behavior of little cats. Instead, the peasant mice seem to have stuck to their view that a cat is a cat, only some cats are bigger and more dangerous than others." Rumors, too, are somewhat subject to reality constraints (Chapter 7).

- rumors in Moscow in the 1930s that gangs of former kulaks threw rubbish, nails, wire, and broken glass into the food to cripple workers who ate it;
- American rumors in September 1942 that crab meat packed by the Japanese contained ground glass.

The coexistence of optimistic and pessimistic rumors with regard to *the same event* – the return of Napoleon I, the fall of Napoleon III – is striking. It is worth asking whether these might mutually fuel each other.

Although the impression from these enumerations that pessimistic rumors are the more frequent could be misleading, a systematic analysis confirms this bias. A study of 1,089 war-related rumors gathered in the United States in September 1942 found that 65 percent had their origin in anger, 25 percent in fear, and only 2 percent in hope. Rumors in the first category did not target the enemy, but allies or domestic groups. They included the rumors that Churchill blackmailed Roosevelt into provoking a war with Japan, that the British sabotaged their own ships in American ports so that they would not have to put out to sea, and that American Catholics were trying to evade the draft.[9] Since both the anger-inspired and the fear-inspired rumors were intrinsically unpleasant, their preponderance is surprising. Why do we so easily believe the worst?[10]

To this quantitative symmetry between optimistic and pessimistic rumors we can add a qualitative asymmetry: rumors inspired by hope are less likely to be used as premises for action. With the important exception of rumors in financial markets, optimistic rumors do not seem to affect behavior. To my knowledge, none of those who hoped for and believed in Napoleon's return left their work to join up with him. By contrast, those who feared and believed in this event took their precautions, such as cutting the grain before it was ripe, hiding their valuables, getting married to avoid being drafted, or withdrawing their deposits from the savings bank. The unfounded rumors of a slave revolt during the War of 1812 caused many members of the militia to flee to protect their homes, thus contributing to the American defeat. In a striking study of the

[9] I do not know whether these rumors arose spontaneously or were encouraged by the enemy. The second idea is consistent with the statement by two social psychologists that "One of the deadliest weapons in the arsenal of psychological warfare is propaganda aimed at convincing some segments of the enemy group that they are suffering more hardships or are gaining fewer benefits than other members of that group."

[10] One author affirms that "more superstitions seem to be associated with bad luck than with good" and suggests that the asymmetry may be "an evolutionary consequence of the need for caution: you're more likely to survive if you can spot potential threats." The asymmetry is indeed puzzling, but I am not sure that the explanation is correct. The tendency to avoid spurious dangers, associated for instance with unlucky numbers, can hardly enhance survival. The author does not discuss the question I raise concerning rumors: are people more likely to *act* on superstitions associated with bad luck than on those associated with good luck?

impact of rumors linking contraceptives with cancer and other serious diseases in the Dominican Republic, unfavorable rumors from trusted sources encouraged many people to discontinue contraception, but rumors favorable to contraception did *not* increase the likelihood using it. I find these two asymmetries puzzling. In individuals, wishful thinking is both common and, when it occurs, often used as a premise for behavior. I do not know why collective wishful thinking tends to be neither.

The dynamics of rumors can illuminate some aspects of this problem. I shall discuss in turn the origin, propagation, and amplification of rumors.

The *origin* of rumors is facilitated by the universal tendency to explain events in terms of intentional agency or objective teleology (Chapter 9). In particular times and places, the formation of rumors can also be facilitated by a preexisting prejudice, schema, or script. In France around 1789, they involved roving bandits, callous grain speculators, a hated seigneurial class, and an oppressive tax collector. A small village may have been more favorable to rumor formation than a larger agglomeration, where skeptics might provide a reality check. Lack of education, too, might favor credulity. In Germany in 1914, the script of *francs-tireurs* in the Franco-Prussian War 1870–1 was easily triggered.

These are, however, only facilitating conditions. It remains to ask, as Lefebvre did: "How did one take the step from suspicion to affirmation?" His answer, which relies on the presence of "some individuals bolder than others," is sketchy, but we can flesh it out with examples from his work. Triggers could be accidental or deliberate. The former included the reflection of a setting sun in a window mistaken for a fire, a misunderstood remark overheard in passing, a joke taken too seriously, and movements of animals mistaken for those of men.[11] The latter included a hawker offering sensationalist rumors to attract buyers, an egocentric trying to make himself important, or somebody just trying to stir up trouble.[12] As these examples indicate, the prospects for a "general theory of rumor formation" are dim. Yet the fact that the Great Fear, with no foundation in reality, erupted *simultaneously and independently* in seven parts of France shows that some ill-understood systemic forces were at work.[13]

[11] In his study about false rumors in World War I, briefer but even more penetrating than Lefebvre's book, Marc Bloch refers to an incident in which the mishearing of the name of the hometown *Bremen* of a German prisoner of war as the name of the French town *Braisne* led to a rumor about a German spy who had been established in France since before the war.

[12] As noted, I am not persuaded by Festinger's claim that people in India came up with rumors to justify their anxieties.

[13] The rumor about Russian troops joining the Allies in August 1914 probably also originated simultaneously and independently in France and England, the only difference being that in France these troops were rumored to be in Marseilles, whereas the English thought they were in Scotland.

A case of a *state-initiated rumor* was the decree issued by the French *contrôleur-général* Orry in 1745, when, as part of a larger statistical effort, he instructed local officials to spread rumors (*semer des bruits*) about an impending tax increase and an impending conscription. The texts to which I have access do not allow me to determine the intention behind this request or its effects. Orry may have been testing the waters, to determine whether these reforms would have been acceptable to the population, since he also asked the officials to assess the resources that might be found to increase the royal revenue. In other words, he might have tried to determine whether these increases were both objectively feasible and subjectively acceptable. Leaders of the Soviet Union, who had an extreme fear of popular unrest, also initiated rumors to test out public reactions to contemplated measures, such as deflation or dismissals.

The *propagation and confirmation* of rumors take many forms. Word of mouth is probably the main mechanism, at least until the arrival of the internet. Montaigne offered what may have been the first analysis of the mechanisms of rumor transmission:

> The distance is greater from nothing to the minutest thing than it is from the minutest thing to the biggest. Now when the first people who drank their fill from the original oddity come to spread their tale abroad, they can tell by the opposition which they arouse what it is that others find it difficult to accept; they then stop up the chinks with some false piece of oakum . . . At first the individual error creates the public one: then, in its turn, the public error creates the individual one. And so it passes from hand to hand, the whole fabric is padded out and reshaped, so that the most far-off witness is better informed about it than the closest one, and the last to be told more convinced than the first. It is a natural progression. For whoever believes anything reckons that it is a work of charity to convince someone else of it; and to do this he is not afraid to add, out of his own invention, whatever his story needs to overcome the resistance.

This mechanism turns on the need for the rumormonger to preempt skepticism by filling in missing details of the story. In other cases, the recipient of rumor may be reluctant to *express* skepticism. Some individuals may be motivated to accept accounts of the danger lest they be accused of cowardice. To display incredulity during the Great Fear was both to invite accusations of serving the counterrevolution by putting the people to sleep in the midst of danger and to risk offending the amour-propre of those who called the alarm. Concerning the rumors of Belgian *francs-tireurs*, their accuracy appeared incontrovertible once they had been used as a premise for bloody reprisals. How else, in fact, could the German atrocities be justified? (Recall Seneca: "Those whom they injure, they also hate.") And when rumors were passed on by wounded soldiers sent home from the front, who would dare to contradict them?

Montaigne's argument assumes that *belief* in a rumor is a condition for passing it on. Field observations and experiments suggest, however, that this

need not be the case. In a study of the transmission of rumors in a neighborhood community, it was found that five of seven individuals who believed the rumor relayed it, and that nine of twenty-three who did not believe it also passed it on. In an experimental study of "revenge rumors" in the workplace, it was found that more credible rumors were not transmitted more frequently. When reflecting on the rationality of rumors, therefore, we should distinguish between the rationality of the belief content of the rumor and the rationality of the strategic use of rumors. In American presidential campaigns, cooked-up rumors to destroy an opponent are common. Franklin Delano Roosevelt survived the rumor that he was suffering not from infantile paralysis, but from syphilis, and Barack Obama the rumor that he was a Muslim or not an American citizen. However, the "swiftboating" rumor may have contributed to the defeat of John Kerry, and Michael Dukakis may have lost the election partly from the rumor that he had seen a psychotherapist (Ronald Reagan referred to him as "an invalid"). While those who transmitted these rumors may not have believed them, they certainly tried to make the "end users" – the electorate – not only believe them, but use them as premises for their votes.[14]

Lefebvre summarized part of his explanation of the Great Fear by saying that "the people scared itself" (*le peuple se faisait peur à lui-meme*). The rumor that brigands were approaching caused the mobilization of troops, which other peasants from a distance mistook for brigands. When a village rang the church bells, the detachments sent out by neighboring villages were mistaken for enemies. In 1848, a warning shot of a cannon in one French village was interpreted in neighboring villages as the din of battle. When rumors of an impending invasion of French paupers reached Germany in March 1848, road workers on the French side of the Rhine crossed the river in a hurry to return to their homes and families. Others, watching from a distance, may have thought they were the French approaching.

The *amplification* of rumors is well documented. Hans Christian Andersen's story of how one feather turned into five hens does not exaggerate the magnifying effect of rumor. After the insurrection of workers in Paris in June 1848, two men who were observed sitting by the side of a country road in Normandy became ten, three hundred, six hundred in the telling and retelling, until finally one could hear that three thousand "levelers" (*partageux*) were looting, burning, and massacring. Thirty thousand soldiers were sent out to counter the threat. An investigation revealed that one of the two was mentally ill and that the other was his father, who was in charge of him. In the same period, a peasant invented a fantasy to scare a child; soon thereafter more than

[14] I do not know of a single instance in which a candidate's handlers have tried to generate a *positive* rumor about him or her. Believing the worst without evidence seems to come more naturally than believing the best without evidence.

a thousand men were in arms to defeat the non-existent "brigands." One mechanism behind such exaggerations is the widespread tendency to *dramatize*: of two narratives that fit the facts more or less equally well, many people will choose the one that is highest in drama content. Another is that nuances in probability assessments and danger assessments easily get lost in the retelling and perhaps that people round up rather than down.

### Informational cascades

Rumors can also arise by entirely rational belief formation, through a mechanism known as "informational cascades." Suppose that each individual in a group has access to some private information about some matter. All form their beliefs sequentially, each one relying on his or her private information *and* on the beliefs expressed by their predecessors (if any) in the sequence. Each villager, for instance, might have some private evidence about the presence of brigands in the vicinity and use it, together with what he has heard from others, to form the opinion he then passes on. Voting by roll call can have the same dynamic, if the matter at hand turns only on beliefs about factual issues and not on preferences. Each member of an assembly will rely not only on her own information but also on what is revealed by the vote of those who precede her in the roll call. For a third example, consider a journal referee for a paper who learns that (but not why) a referee for another journal had favored rejection.

In these situations people use the conclusions of the belief formation process of others as indirect inputs to their own belief formation, without knowing the direct inputs (the private information) others used to form their conclusions. It may then happen that rational individuals end up with false beliefs, although they would have reached the correct conclusion had each of them had access to the "raw data" of their predecessors and not only to their conclusions. In the referee example, the second reader might, if he had read the first report, have spotted bias or faulty reasoning. Yet if he only knows the conclusion of the first report and the fact that the first journal is highly respectable, he should rationally take into account the negative opinion of the first referee together with his own assessment. If the latter is favorable but only slightly so, he may end up recommending rejection. A third reviewer with a strongly favorable personal opinion might also (rationally) favor rejection if she learns that two previous referees did so. Yet the outcome may be suboptimal (relative to the aims of the scholarly community), since the second and third referees favored publication and the first may have been only mildly against it.[15] If the

[15] Various conformity-reducing practices may be understood in this perspective. When (as used to be the case in Norway) the internal grader at university exams sends the student's answer to an

reviewers had read the article in the inverse order, the conclusion would have been different ("path-dependence").

### Bibliographical note

The line-matching experiments, first carried out by Solomon Asch, are described in any textbook on social psychology, for example, E. Aronson, *The Social Animal*, 9th edn (New York: Freeman, 2003). The moving-light experiment is described in R. C. Jacobs and D. T. Campbell, "The perpetuation of an arbitrary tradition through several generations of laboratory microculture," *Journal of Abnormal and Social Psychology* 62 (1961), 649–58. The data on Israeli and Palestinian preferences for a return to the 1967 borders are from J. Shamir and M. Shamir, "Pluralistic ignorance across issues and over time," *Public Opinion Quarterly* 61 (1997), 227–60. The reference to the pressures on Pasternak is from S. Fitzpatrick, *Everyday Stalinism* (University of Chicago Press, 1999), p. 198. For drinking on campus, see D. A. Prentice and D. T. Miller, "Pluralistic ignorance and alcohol use on campus: some consequences of misperceiving the social norm," *Journal of Personality and Social Psychology* 64 (1993), 243–56. For the students exposed to the incomprehensible article, see D. T. Miller and C. McFarland, "Pluralistic ignorance: when similarity is interpreted as dissimilarity," *Journal of Personality and Social Psychology* 53 (1987), 298–305. The unraveling scenario relies on T. Kuran, *Private Truths, Public Lies* (Cambridge, MA: Harvard University Press, 1995). The observation about the effect of the Reformation on the belief in the king's power to heal is due to M. Bloch, *Les rois thaumaturges* (Paris: Armand Colin, 1961). The letter (by Hannah Greg) citing the effects of "fear acting upon prejudice" in rumor formation is cited from J. Uglow, *In These Times* (London: Faber and Faber, 2014), p. 229. The main studies of rumor formation on which I have drawn are G. Lefebvre, *La grande peur de 1789* (Paris: Armand Colin, 1988); F. Ploux, *De bouche à oreille: naissance et propagation des rumeurs dans la France du XIXe siècle* (Paris: Aubier, 2003); R. Cenevali, "The 'false French alarm': revolutionary panic in Beden, 1848," *Central European History* 18 (1985), 119–42; M. Bloch, "Réflexions d'un historien sur les fausses nouvelles de guerre," *Revue de synthèse historique* 33 (1921), 13–35; and C. Prochasson and A. Rasmussen (eds.), *Vrai et faux dans la Grande Guerre* (Paris: Éditions La Découverte, 2004). The rumors among the Galician peasantry are mentioned in L. Namier, *1848: The Revolution of the Intellectuals* (Oxford University Press, 1946), p. 15. A detailed catalogue and analysis of rumors in ethnic riots are found in D. Horowitz, *The Deadly*

external grader, he does not pass his own grade along. For the same reason, when seeking a second medical opinion, one should not tell the second doctor what the first said.

*Ethnic Riot* (Berkeley: University of California Press, 2001), pp. 76–8. For rumors in the stock market, see A. M. Rose, "Rumor in the stock market," *Public Opinion Quarterly* 15 (1951), 61–86. The study of rumors during World War II is R. Knapp, "A psychology of rumor," *Public Opinion Quarterly* 8 (1944), 22–37. The asymmetry between "bad-luck" and "good-luck" superstitions is noted in D. Hand, *The Improbability Principle* (New York: Scientific American, 2014), p. 248. The claim that propaganda may aim at convincing part of the enemy population that it is treated worse than others is from D. Krech and R. Crutchfield, *Theory and Problems of Social Psychology* (New York: McGraw-Hill, 1948), p. 411, cited after F. Heider, *The Psychology of Interpersonal Relations* (Hillsdale NJ: Lawrence Erlbaum, 1958), p. 289. The rumor about kulaks putting broken glass in the food of workers is reported in S. Fitzpatrick, *Everyday Stalinism* (University of Chicago Press, 1999), p. 45. An exceptionally detailed catalogue of rumors in India following the 1934 earthquake is J. Prasad, "A comparative study of rumours and reports on earthquakes," *British Journal of Psychology* 41 (1950), 129–44. The comments on state-initiated rumors in the Soviet Union draw on the entry "Rumors" in I. Zempsov, *An Encyclopedia of Soviet Life* (New Brunswick, NJ: Transaction Publishers, 2001). The comments on the French decree of 1745 draw on F. de Dainville, "Un dénombrement inédit au XVIIIe siècle: l'enquête du Contrôleur général Orry – 1745," *Population* 7 (1952), 49–68. The study of the transmission of beliefs in a neighborhood community is L. Festinger *et al.*, "A study of a rumor: its origin and spread," *Human Relations* 1 (1948), 464–86. The study of revenge rumors is P. Bordia *et al.*, "Rumor as revenge in the workplace," *Group & Organization Management* 39 (2104), 363–88. An introduction to the mechanism of informational cascades is S. Bikchandani, D. Hirshleifer, and I. Welch, "Learning from the behavior of others: conformity, fads, and informational cascades," *Journal of Economic Perspectives* 12 (1998), 151–70.

# 23 Collective action

Collective action can be defined as decentralized action by the members of a group to eliminate public bads that affect all of them or to create public goods that benefit all of them. Although such actions benefit everybody (they are *Pareto-superior* to the status quo), it may not be in the self-interest of any individual to participate or contribute. To show the pervasive character of this dilemma between the individual and the collective good, I begin the chapter by offering a large number of examples, before discussing their fine grain.

In some cases, successful collective action arises out of horizontal interaction among group members. In other cases, a government or an institution may promote Pareto-improving actions by vertical measures (rewards or punishments) that affect the choice set, and therefore the choices, of the members. Not all vertical measures generate Pareto improvements, however; some create winners and losers. In the latter case, the obstacle to their implementation is group interest, not individual self-interest. In the next two chapters I consider several cases of that kind.

## Some collective action problems

A collective action problem is a many-person Prisoner's Dilemma (Chapter 18), at least as far as material incentives are concerned. To illustrate the variety and ubiquity of collective action, I shall present a number of examples, several of them mentioned in earlier chapters or discussed later. In enumerating the cases, I identify them by the cooperative strategy.[1]

- Voting in elections.
- Elected members of parliament being present for votes and debates.

[1] In spite of the intuitive positive connotations of the word, "cooperation" is used here as a technical term, with no implication for the desirability or social utility of the action. Cartels are almost universally, and trade unions occasionally, viewed as acting contrary to the general interest. Criminal gangs have collective action problems arising from snitching.

- Committee members or market traders taking the time to inform themselves rather than counting on someone else to gather information ("informational free riding").
- Joining an insurrectional movement.
- Joining a trade union.
- In revolutionary America, joining a non-importation movement or a non-consumption movement.
- Industry unions making moderate demands for wage increases. (An encompassing trade union, organizing workers in all industries, will pull its punches out of self-interest, since all members will feel the full impact of the price increase triggered by high wages. Workers in one industry, however, usually do not consume so much (if anything) of its output that the price increase offsets their wage increase.)
- Respecting an agreed-upon cartel prize or volume of production.
- Abstaining from littering, polluting, overgrazing, overfishing.
- Refusing government subsidies. "Democratic centuries are times of trial, innovation, and adventure. There are always a host of men engaged in difficult or novel enterprises which they pursue independently, unencumbered by their fellow men. These people accept the general principle that the public authorities should not intervene in public affairs, but each of them seeks, as an exception to this rule, help in the affair that is of special concern to him and tries to interest the government in acting in that area while continuing to ask that its action in other areas be restricted. Because so many men take this particular view of so many different objectives at the same time, the sphere of the central power imperceptibly expands in every direction, even though each of them wishes to restrict it" (Tocqueville). Marx made a similar claim: "The attitude of the bourgeois to the institutions of his regime is like that of the Jew to the law; he evades them whenever it is possible to do so in each individual case, but he wants everybody else to observe them."
- Doing military service. Free-riding tactics for avoiding military service include cutting off a finger, paying a substitute, taking a debilitating drug prior to the medical examination, obtaining a needless orthodontic treatment, getting married, going to college, going into hiding, or leaving the country.
- Reporting one's income correctly and paying one's taxes on time.
- In federal systems, each state contributing its fair share of taxes and soldiers. This problem also arose in France under the *ancien régime*, when the country was divided into three *estates* (clergy, nobility, commoners), which constantly fought to achieve tax exemptions.
- In military systems, each branch (air force, navy, army) providing accurate rather than self-serving estimates of the best way to conduct or prepare for war.

- Shipowners contributing to build a lighthouse.
- Car drivers abstaining from using a new road connecting two cities. *Braess's paradox* shows that the result of making a new road available can be to increase the average time of driving, for a given number of cars.[2] In Stuttgart, after investments in the road network in 1969, the traffic situation did not improve until a section of newly built road was closed for traffic. In 1990 the closing of 42nd Street in New York City reduced the amount of congestion in the area (Wikipedia).
- In classical Athens, rich citizens voluntarily providing public goods for the city (evergetism). In the words of its foremost historian, "It was ... in the interest of each notable not to sacrifice himself to the ideal, and to let others shine in his stead." If they nonetheless stood fast, it was because of "the precise fear of vague sanctions and vague fear of precise sanctions."
- Feudal lords staying at court rather than at their estates. In England, in one view, "the barons regarded this attendance as their principal *privilege;* in another, as a grievous *burden*. That no momentous affairs could be transacted without their consent and advice, was in *general* esteemed the great security of their possessions and dignities: but as they reaped no immediate profit from their attendance at court, and were exposed to great inconvenience and charge by an absence from their own estates, every one was glad to exempt himself from each *particular* exertion of this power; and was pleased both that the call for that duty should seldom return upon him, and that others should undergo the burden in his stead" (Hume). In his book on the *ancien régime*, Tocqueville made the opposite argument with regard to France: the nobility as a class would have benefited if the nobles had stayed on their estates, but the individual noble preferred the lures of the court. "The limbs gained at the expense of the body."
- Popes showing intertemporal solidarity with other popes and monarchs international solidarity with other monarchs. Concerning popes, Hume writes that "The industry and perseverance are surprising, with which the popes had been treasuring up powers and pretensions during so many ages of ignorance; while each pontiff employed every fraud for advancing purposes of imaginary piety, and cherished all claims which might turn to the advantage of his successors, though he himself could not expect ever to reap any benefit from them." Concerning monarchs, Queen Elizabeth I refused to support the Dutch revolt against her enemy Philip II of Spain because, as a reigning monarch she "view[ed] with suspicion those whose claims arguably constituted an encroachment on the rights of sovereigns."

[2] A better-known problem (discussed in Chapter 17) is that the new road may *encourage more drivers* to use their cars.

- Families limiting their number of children.
- Participating in community work.
- Soldiers using latrines when available. A description of what happens when soldiers instead defecate at their convenience can be taken from a book about the Italian army in World War I: "Incredibly, the men could not see how needlessly unpleasant they made life for everybody – themselves included – by not using the latrines. This chronic inability to grasp the wider effect of their actions was a trait that [an observer] dubbed 'the *cretinous egotism* of the Italian.'"
- Letting one's children be vaccinated. This constitutes a collective action problem when two conditions are satisfied: (i) vaccination has a small risk of serious side effects; (ii) if nearly everyone is vaccinated, they provide protection for the non-vaccinated.
- Doctors abstaining from treating ordinary infections with antibiotics, whether they are viral or bacterial. This is not the only mechanism by which overuse of antibiotics generates a collective action problem. Other mechanisms include excessive prophylactic use of antibiotics by doctors, the use of wide-spectrum drugs when a narrow-spectrum drug would suffice, and the use of antibiotics to enhance growth in livestock.
- Nations or (in federal systems) states resisting the temptation to offer low corporate taxes, to prevent a "race to the bottom."
- Ignoring social norms of wasteful conspicuous consumption, such as SUVs.
- Making charitable donations (see Chapter 5).
- Reducing water consumption in times of shortage (see Chapter 5).

### The technology of collective action

In the following I shall assume that no actor is "big enough" to have a selfish reason for cooperating even when no one else does so. By making this assumption, I exclude two interesting cases. First, when there is *one* big actor and many small ones, for instance one large shipowner and many small ones, the latter may abstain from making a financial contribution to the building of a lighthouse if they expect that the former will do so out of his self-interest. If one trade union organizes a majority of the workers, while the rest are organized in many small unions, the former may make moderate wage demands out of self-interest, while none of the latter has any incentive to restrain itself. This phenomenon has been called "the exploitation of the large by the small." Second, if there are *two* big actors, their relation may be that of a Game of Chicken (Chapter 18): each shipowner will build the lighthouse on his own if and only if he is certain that the other will not. The union case is different: each of two large trade unions might have an interest in making moderate demands, regardless of what the other does,

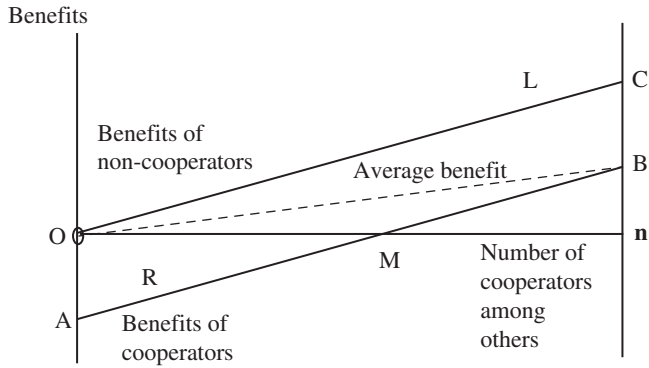



Figure 23.1 Redrawn from T. Schelling, *Micromotives and Macrobehavior* (New York: Norton, 1978)

if its members as consumers would be sufficiently hurt by the inflation caused by high demands.

In a group of n + 1 individuals, Figure 23.1 indicates how the pay-off to a given individual varies as a function of his own behavior and of that of the others.[3] The behavior of others is indicated along the horizontal axis, which measures the number of cooperators (among these others). If the individual is also a cooperator, his utility, measured along the vertical axis, is indicated along the R line AB in the diagrams. If he is a non-cooperator, his utility is measured along the L line OC. The L and R lines intersect the vertical axes in the order that defines the ordinary (two-person) PD: the most preferred outcome is unilateral non-cooperation (free riding), the next best is universal cooperation, the third best universal non-cooperation, and the worst outcome unilateral cooperation (being exploited). As in the two-person case, non-cooperation is a dominant strategy, since the L line is everywhere above the R line. In contrast to the two-person case, however, we can define a number of cooperators M that can make themselves better off by cooperating, even in the presence of free riders whom they make even better off. The line OB shows the *average* benefit to everybody, cooperators and non-cooperators, as a function of the number of cooperators. As the number of agents is constant, OB will also reflect the *total* benefit produced by cooperation.

[3] I shall write as if individual choices are binary (individuals either cooperate or they do not) and as if outcomes are continuously variable (a public good, such as clean air, may be provided to a smaller or larger extent). In reality, individuals may differ in *how much* they contribute, not merely in *whether* they do. I shall not take this complication into account. Also, some public goods are "lumpy" or discrete. If individuals are in a community lobby to keep the local school open, it will either close or remain open. This complication can be finessed by interpreting the outcome as the continuously variable *probability* of the public good's being provided.

The situation in Figure 23.1 reflects a special case. It assumes that the cost of cooperation, measured by the distance between the L and R curves, is constant. In other cases, the cost of cooperation increases as more people cooperate. As people join call-in campaigns for public radio, the lines become congested and it takes more time to get through. It may then happen that the last to join[4] actually reduce the average benefit, because the cost to them of participating exceeds the sum of the benefits they generate for everybody else (and for themselves). The cost may also be high initially and then decrease. As more people join up in a demonstration, policy and security forces have to spread themselves more thinly, unless new troops can be called in.

Figure 23.1 also assumes that the benefits of cooperation, given by the L line, are a linear function of the number of cooperators. Each new cooperator adds the same amount to everybody's welfare. Increasing marginal benefits can be illustrated by cleaning a beach of litter: the last bottle that is removed makes more of an aesthetic difference than the penultimate one. Decreasing marginal benefits are also frequent. A simple example is calling city hall about a pothole in a middle-class urban area: the first person who takes the time to call could make the probability 0.4 that the hole will be fixed, the second raise it to 0.7, the third to 0.8, the fourth to 0.85, the fifth to 0.88, and so on. Sometimes, both the first and last contributors add little, whereas those in the middle are more efficacious. A few revolutionaries or strikers do not make much of a difference, and when almost everyone has joined it matters little whether the few uncommitted do so too. In social movements, this pattern is probably typical.

The marginal benefits of cooperation may even be negative over some range of cooperators. Unilateral disarmament can make all nations worse off if it creates a power vacuum to be invaded, thus unleashing a general war. Isolated acts of rebellion may give the authorities a pretext for cracking down on potential rebels as well as on the actual ones. Conversely, there may be too many cooperators. Suppose that in wartime everybody insists on joining the army, so that industries vital to the war effort are understaffed and the war is lost. If everyone insists on helping out with the dinner at the outing, the many cooks may spoil the broth.

As these remarks show, the technology of collective action differs from case to case. In the following, I focus on the case shown in Figure 23.2, which I believe to be fairly typical of social movements trying to bring about a

[4] Here and elsewhere, words such as "first," "middle," and "last" can refer to the times at which successive cooperators join, as in building a revolutionary movement. But they can also refer to simultaneous acts of cooperation, as in voting. To say that the last voters add little is to say that the benefit created in a situation in which everyone votes is nearly the same as the benefit created when almost everyone votes.

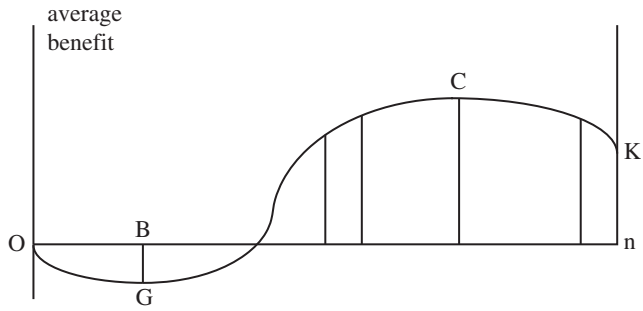



Figure 23.2

change of policy. The first contributors incur high costs or risks and produce few benefits for others. They may, in fact, harm others rather than benefit them. Their net contribution is negative. The last contributors also produce few benefits. In some cases, their cost may be decreasing, as I suggested. In other cases, all who fight for a cause may incur considerable costs or risks until the adversary capitulates. Those who joined the French resistance in 1944 often did little damage to the Germans, but all ran a considerable risk to their lives.

### Unraveling

Cooperation can unravel in a variety of ways. As an illustration, I offer the following fable in which members of university departments or workers' collectives may recognize themselves. Initially, the group handles its common affairs informally, on the basis of trust and cooperation. At some point, a *black sheep* makes his or her entrance, and the group is transformed. The motives of appointed or elected leaders are questioned. Procedure takes precedence over substance. Many low-level decisions are appealed to higher authorities. The decisions and internal workings of the organization are denounced in the social media. When allocating benefits such as travel allowances among members of the group, the norm is "Nobody shall have what not everyone can have." Trust is replaced by suspicion, and cooperation by opportunism or worse. Everything takes twice as long, and results are worse. While not reflecting a universal tendency, the fable is more than anecdotal.

The following public good experiment (a stylized version of many actual experiments) offers a formal demonstration of unraveling. Each experiment has several rounds, in each of which the participants receive 10 monetary units (MU), which can be converted into cash at the end of the session. They can either keep the tokens to themselves or donate them to a common pool. If they donate, the donations are multiplied, but not enough to make it selfishly

Table 23.1

| | | Subjects | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | Mean |
| Periods | 1 | 10 | 10 | 10 | 0 | 7.5 |
| | 2 | 7.5 | 7.5 | 7.5 | 0 | 5.6 |
| | 3 | 5.6 | 5.6 | 5.6 | 0 | 4.2 |
| | 4 | 4.2 | 4.2 | 4.2 | 0 | 3.1 |
| | 5 | 3.1 | 3.1 | 3.1 | 0 | 2.4 |
| | . | | | | | |
| | . | | | | | |
| | 10 | 0.4 | 0.4 | 0.4 | 0 | 0.3 |

rational to donate. If all donate, however, all gain. The structure, in other words, is that of a many-person Prisoner's Dilemma. After each round, the participants are told how many donated and how much, before deciding whether and how much to donate in the following round. In other words, they make their decisions on the basis of their knowledge about outcomes, not actions. They are not told *who* donated and who did not, nor, if they had been told, would they have any way of taking account of this information by punishing specific individuals.

Suppose there are four participants and that the game has ten rounds. Three participants are *perfect conditional cooperators*, in the sense that they initially donate all their MU and, in later rounds, donate the average of the donations in the previous round. The fourth is a *perfect egoist*, who prefers to free ride and never donates anything. In one sequence, the game could go as shown in Table 23.1.

After ten rounds, the initially high contributions have fallen almost to zero. If the fourth agent had also been a perfect conditional cooperator, no unraveling would have occurred. It is the one black sheep that spoils the flock. If the proverb is right and every flock has a black sheep, the prospects for sustained cooperation are poor.

Prospects may be improved, however, if the group contains an unconditional cooperator. In the hypothetical experiment shown in Table 23.2, the third subject always donates 10 MU, regardless of what others do. With this single "white sheep," the amount donated by each conditional cooperator ("grey sheep") converges to 5 MU per round. Overall, the donations amount to 20 MU per round, half of what would have been achieved with four perfect cooperators, but much more than the outcome of the game in Table 23.1. To a considerable extent, the white sheep neutralizes the black sheep.

Outside the laboratory, the decline in trade union membership in some countries may be due to an unraveling mechanism. In many American states,

Table 23.2

| | | Subjects | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | Mean |
| Periods | 1 | 10 | 10 | 10 | 0 | 7.5 |
| | 2 | 7.5 | 7.5 | 10 | 0 | 6.2 |
| | 3 | 6.2 | 6.2 | 10 | 0 | 5.6 |
| | 4 | 5.6 | 5.6 | 10 | 0 | 5.3 |
| | 5 | 5.3 | 5.3 | 10 | 0 | 5.1 |
| | . | | | | | |
| | . | | | | | |
| | 10 | 5 | 5 | 10 | 0 | 5 |

"right-to-work" legislation prevents unions from requiring non-unionized workers to join the union ("closed shop") or, if they do not join, to pay a negotiating fee to the union. Without a "fair-share" provision of this kind, workers may prefer to be free riders on the union effort to negotiate wage increases. Although the importance of this effect is debated, common sense and much scholarship suggests that it is non-negligible.

### Maintaining cooperation

Given the ubiquitous free-rider temptation, how do we explain the substantial amounts of cooperation that we observe? In addition to the presence of unconditional cooperators, cooperation by selfish individuals may be sustained by selective *rewards* and selective *punishments* (see Chapter 18 and Chapter 25). Here, rewards and punishments are taken in a wide sense, which includes any deliberate action to lower the costs of cooperation and increase the costs of non-cooperation *by any means*, for instance by acting on prices. Rewards as well as punishments can be provided either *vertically*, by the state or an organization, or *horizontally*, by other members of the group that faces a collective action problem. As horizontally provided rewards seem marginal, I shall ignore them. I am not denying, of course, that people can be motivated by the desire to be praised by their peers, only that this desire is an important factor in overcoming the free-rider problem. Although some laboratory experiments find this effect, I doubt they matter "in the wild." We do not praise people for voting, putting the ice cream wrapper in a garbage can, or vaccinating their children.

*Rewards* can be important in sustaining unions, since members can get access to the union's summer camp, cheaper insurance than individuals could negotiate on their own, and other goods. The collective action problem in voting can be alleviated by paying voters for showing up, as was done in

classical Athens.[5] In November 2006, a ballot initiative in Arizona to award $1 million to a randomly selected voter, for the purpose of increasing turnout, was rejected by roughly one million votes against and half a million in favor. The Norwegian municipality Evenes successfully adopted a similar policy in 1995, when turnout went up from 63 percent to 71 percent after it set up an election lottery with a trip to southern Europe as a prize. In 2009, the Norwegian municipality of Høyland offered a prize of 100,000 Norwegian crowns, about $20,000, to the electoral district with the highest turnout.[6] The winner, Utsira municipality, with a total population of 216, achieved a turnout of 92.5 percent. As the mayor observed, the distance to the polling station was short; as he did not say, the ease of detecting and ostracizing non-voters probably also contributed to the high turnout. This kind of collective reward system is similar to the use of team bonuses, which are also more likely to make people cooperate if the group is small enough that members can monitor each other.[7]

*Punishment* is widely used to deter non-cooperative behavior. Vertical punishment is based on the law or on the by-laws of organizations, which can fine or exclude their members. Parliament could, but rarely does, fine members if they are absent without a good reason. When people pay their taxes, and pay them on time, it is at least in part because of the fines and jail sentences that can be imposed on tax evaders or procrastinators. In some countries, voting is mandatory. Vaccination, too, is often a duty rather than a right. China's one-child policy was imposed to block the incentive of families to have many children who will take care of them when they grow old. The use of polluting products can be limited by taxing them. Since readers can easily multiply examples, I turn to the more complex case of horizontal punishment.

Many experiments have found that the punishment of non-cooperators by co-subjects is effective in preventing the unraveling of cooperation. In addition to (i) keeping his endowment or (ii) donating some of it to the common pool, a subject may (iii) use part of it to punish another subject, usually for being

[5] To encourage the voting of those who have some, although not a very strong, sense of civic duty, the state can lower the cost of voting by making registration easier and providing many polling stations.

[6] To increase its own chances of winning, it offered a free pizza to all voters, in an amusing contrast to earlier centuries when a candidate would ply voters with food and drink to make them vote *for him*.

[7] Jejune economists and political scientists have tried to find other private benefits for participants in collective action. Thus voting has been explained by its psychic benefits, or the warm glow from doing one's duty. As Kant noted, this hypothesis can never be definitely excluded, but that fact is hardly a reason for adopting it. As I discussed in Chapter 5, the warm-glow theory has also been invoked to explain charitable donations; as I argued there, the merits of the explanation are doubtful. Finally, some scholars have asserted that people join revolutionary movements because they are motivated by the private reward of becoming leaders in the post-revolutionary order, rather than by the collective good of a better society. In the Conclusion, I argue that this claim is unproven and implausible.

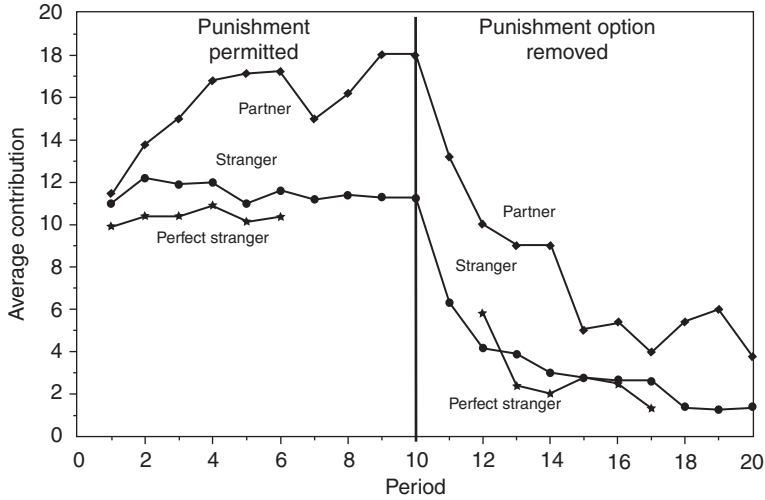



Figure 23.3

non-cooperative.[8] The punishment takes the form of reducing the endowment of the target person, usually by less than what it costs to punish him. In the representative public goods game showed in Figure 23.3, subjects interacted with an anonymous partner from a previous round, with randomly assigned individuals (strangers), or with randomly assigned individuals with whom they would never interact more than once (perfect strangers). Allowing for the punishment of non-cooperators had a strong effect on the level of cooperation – strongest, perhaps surprisingly, on the partners. Earlier (Chapter 21) I suggested that the target of punishment may be more affected by how much it costs *the punisher* than by what it costs him.[9] To my knowledge, no experiment has been run that would allow us to distinguish the effects of these two variables.[10]

[8] Some intriguing experiments show, however, that subjects sometimes punish *cooperators*, perhaps for being "do-gooders" who make others feel ashamed.

[9] Experiments show that unraveling may also be prevented if members of the group can award "disapproval points" to non-cooperators. Although the effect is less than when they can award monetary "punishment points," the difference may be due to the fact that awarding disapproval points is costless.

[10] To do so, one would have to compare the effect of one set of vectors (p, q) of costs to the punisher and costs to the target person with the effect of another set $(p + \partial, q)$, $\partial > 0$. Because a punisher will usually punish less if it costs him more, one would have to tell him that the cost was p while telling the target person, falsely, that they were $p + \partial$. Experimenters are reluctant, however, to lie to their subjects about the pay-off conditions, since if the practice became known the credibility of the instructions in later experiments would be undermined.

What is the relevance of these findings "in the wild," outside the laboratory? In everyday life, we do not go around punishing people who misbehave merely by lack of cooperativeness.[11] Instead we express *disapproval*, avoid them if we can, and keep our distance if we cannot.[12] In the laboratory, how strongly I punish misbehavior depends on three factors: how badly I think the other person behaved, the cost to me of punishing him, and the cost to him of being punished. In ordinary interactions, how strongly I express disapproval depends mainly on the first factor, although instrumental considerations can also matter. Moreover, what makes disapproval for the target so painful is the face-to-face character of the interaction, in contrast to the artificially anonymous conditions in the laboratory. There is no doubt that anticipation of face-to-face disapproval of free riding can prevent cooperation from unraveling, as I suggested in the comments on collective reward systems and team bonuses. The question, to which I have no answer, is whether the experiments demonstrate a punishing propensity that is *reinforced* by personal interactions or whether interactions *preempt* the triggering of that propensity.

An important collective action problem in modern societies is that of tax payment and tax collection. How can citizens be made to declare their income honestly and pay their taxes on time? Although facts are hard to come by, it is estimated that in the United States 18–19 percent of reportable income is not reported to the Internal Revenue Service, causing an annual revenue loss of $450–500 billion, about four times as much the federal budget for education. In addition to the obvious coercive tools at the disposal of the government, could horizontal mechanisms, such as social norms and quasi-moral norms (Chapter 5) also induce compliance?

Considering first quasi-moral norms, it is often argued that people are more willing to report their taxes correctly if – but only if – they believe that most others are doing so. They are conditional cooperators, wanting to do their "fair share," but not willing to be taken advantage of. This may well be true, but where do these beliefs come from and how accurate are they? We may compare tax compliance with the effort to reduce water consumption in Bogotá that I mentioned in Chapter 5. In that case, the citizens had access in real time to data about the aggregate consumption. In the American Revolution, "newspapers made it possible for Americans to imagine that virtual strangers were

[11] Adam Smith insisted on this point: "Though Nature . . . exhorts mankind to acts of beneficence, by the pleasing consciousness of deserved reward, she has not thought it necessary to guard and enforce the practice of it by the terrors of merited punishment in case it should be neglected." He also asserted that lack of gratitude cannot be punished. Seneca claimed, however, that the Macedonians had laws against ingratitude, a claim that has also been made about ancient Persia.

[12] If those who misbehave are powerful, we might, Seneca suggests, avoid showing that we avoid them. Gibbon observed that it was dangerous to trust the Emperor Augustus, and even more dangerous to show one's distrust.

actively supporting each other" in the non-importation campaign. With regard to underreported income, there are no comparable data. It is highly unlikely that citizens read technical econometric studies of the incidence of tax evasion before deciding whether it is low enough to keep them honest. In Norway, as I mentioned in Chapter 21, citizens can access the internet to peek at the income and tax data of their friends or neighbors, and compare them with their observed lifestyle. Also, multimillionaires who pay little or no income tax are regularly denounced in the newspapers. Such information does not, however, provide the basis for a rational belief about the average level of compliance. Nor could any untrained person – and probably no trained person either – draw inferences about tax compliance from data about the provision of the public goods that are funded by taxes.

*Social* norms may be more effective in sustaining voting as well as honest reporting of income. In Argentina, people who might be tempted to stay home on election day could be deterred by the fear that neighbors and friends could ask them, "So, I noticed that you didn't vote – how come?" In Norway, people who might be tempted to cheat on their taxes could be deterred by the fear of intrusive questioning – "How could you afford that expensive car, after buying that new cottage last year?" In addition, they might fear being denounced to the tax authorities. If the internet access scheme was set up to instill this fear, it seems to have been successful.

### Snowballing

In previous sections I have discussed the unraveling of collective action, and some mechanisms that can prevent unraveling. Here I discuss the process of *snowballing from non-cooperation to cooperation*, using as an example the non-violent revolutionary collective action that took place throughout Eastern Europe in the summer and fall of 1989. I shall consider both *within-country* and *between-country* snowballing. Although the latter was not itself a form of collective action, it may have been a facilitator of the within-country processes.

I take it for granted that the situation in the Communist bloc before 1989 was suboptimal for everybody except a small elite, and that collective action was needed to reach a better equilibrium. Unlike countries occupied by Nazi Germany, the satellite Communist countries never saw any assassinations of domestic or foreign oppressors, perhaps because there were no parallels to the organized resistance organizations in Western Europe during World War II or perhaps because potential assassins thought the population would blame them rather than the regime for the inevitable reprisals. Opposition to the regimes mainly took the form of public demonstrations. My focus here is on the snowballing of these demonstrations. The most striking examples were the demonstrations in Leipzig on successive Mondays in the fall of 1989, with

numbers increasing from 8,000 (the largest since the insurrection in 1953) on September 25, through 20,000 on October 2, to 70,000 on October 9. Even though the regime was widely detested or hated, that fact by itself was not what brought people out onto the streets. A similar scenario had been enacted earlier in Budapest, and would be enacted later in Prague.

As I noted earlier, Poland in early 1989 was characterized by a state of pluralistic ignorance. Foreign observers, members of the government and of the Solidarity movement, and presumably the population at large, believed that enough people would vote for the Communists in the first semi-free elections to ensure a Communist majority in parliament (*Radio Free Europe Research,* April 7, 1989). They were proven wrong: Solidarity swept the free part of the June elections. The reason why the Solidarity negotiators in the Round Table Talks only made (what they wrongly believed to be) relatively modest demands was the fear that a radical regime change might trigger Soviet intervention. When the demands did lead to a radical change (because of the pluralistic ignorance) and the Soviets did not intervene, other countries, notably Hungary, could use rough Bayesian updating (Chapter 13) to revise their beliefs about this risk.

It may be useful to distinguish between *political contagion* and *political snowballing* across countries. A classical case of revolutionary contagion occurred in 1848, when the February revolution in France triggered similar regime changes in many other European countries. More recently, the Arab Spring provides an example. The psychological mechanism of contagion remains opaque. It is not clear why information about a revolution in one country should make people take to the streets in another country with different economic, social, and political conditions.[13] After all, no material or cognitive obstacles prevented the Egyptians from doing *before* the revolution in Tunis what they did after it erupted. If events in Tunisia shaped events in Egypt, it was not by virtue of providing new information or by exporting money, arms, leaders, or fighters. The commonly used phrase of "providing inspiration" is vacuous; yet *some* causal mechanism must have been at work.

Political snowballing occurs when events in one country provide information that was previously *not* available to citizens in another. In stylized form, it involves three countries: A, B, and their common hegemon C. The citizens of A and B may be reluctant to rise up against C, fearing a violent oppression. If, nevertheless, there is an uprising in A and C does *not* react, the citizens of B may use Bayesian updating for a downward revision of the probability of C intervening in the case of an uprising in country B. During the Cold War, the "domino theory" stated that if C was America and A and B were two countries

[13] Nor is it clear why the Arab Spring scared the Chinese authorities so much that they banned news about it.

within the American sphere of influence, a failure of America to stop a Communist takeover in A would serve as a green light for insurrection in B. The theory served as the justification for the American war in Vietnam. Although the phrase "domino theory" was to my knowledge not used about the Communist world, the Soviet leadership probably thought in similar terms when they crushed uprisings in Berlin, Budapest, and Prague. In 1989, with Poland and Hungary in the role of A and B and the Soviet Union in the role of C, the non-intervention in Poland probably did serve as a green light for the Hungarian transition, by reducing the expected costs of participation in collective action. Moreover, by some further steps of Bayesian updating, Soviet passivity in Hungary probably reduced the expected costs of participation in East Germany and Czechoslovakia.

While private costs and risks of participation do enter into the decisions of individuals to take to the streets, other motivations matter as well. In classifying them, I shall cite their *double heterogeneity*, qualitative as well as quantitative. Different agents may have qualitatively different motivations. These include the personal costs and risks of participation, consequentialist and non-consequentialist moral norms, social norms, and quasi-moral norms.[14] The triggering of norms may also, as noted earlier, be sensitive to quantitative thresholds. Consequentialist moral norms range from utilitarianism to various degrees of altruism. The more people care about the welfare of others, the more willing they are to suffer risks, and the earlier they will join the demonstrations at a stage when there is less "safety in numbers." With regard to social norms, one person may join a movement if a single friend or neighbor expresses disapproval of his passivity, whereas another may require the sustained pressure of many. With regard to quasi-moral norms, one person may join a march once he observes that a hundred others have done so, whereas another waits until the crowd has grown to a thousand. The triggering of non-consequentialist moral norms is of course threshold-independent. At the other extreme, those who care only about their personal costs and risks will never participate in collective action.

On this background, let me sketch a generic scenario for snowballing:

1. Everyday Kantians, saints, heroes, and the slightly mad initiate collective action without regard to consequences.

[14] As noted earlier it is unlikely that participants in movements for regime change were motivated by the expected material benefits that will accrue to them as leaders of the new regime. Also, some scholars have suggested other motivating factors, such as "process benefits" or "agency benefits," that is the pleasure of participating in a collective movement or of experiencing an enhanced sense of being in control over one's own fate. Whatever the importance of these benefits once a movement is underway, it seems implausible that the *expectation* of receiving them could motivate agents to join.

2. If and when the first group is large enough for individual participation to produce a positive expected benefit, *net of costs and risks to the agent,* people motivated by consequentialist norms come on board.
3. Some people motivated by quasi-moral norms will join when the observable subset of (1) + (2) exceeds a certain numerical threshold p.
4. Some people motivated by social norms will join when the subset of (1) + (2) who observe them exceeds a certain threshold q.
5. Other people motivated by quasi-moral norms will join when the subset they observe of (1) + (2) + (3) + (4) exceeds a threshold $p^* > p$.
6. Other people motivated by social norms will join when the subset of (1) + (2) + (3) + (4) who observe them exceeds a threshold $q^* > q$.
7. And so on.

Depending on the distribution of motivations and of thresholds in the population, the process may not get beyond (1) and run out of steam quickly, or end up reaching everybody.

Historians or other scholars will probably never be able to reconstitute a full process of this kind. The micro-mechanisms can be documented, but their relative importance and the way in which one creates the conditions for the triggering of another cannot be established without evidence they are unlikely to possess.

## Bibliographical note

A classical study of collective action, emphasizing cooperation induced by selective rewards, is M. Olson, *The Logic of Collective Action* (Cambridge MA: Harvard University Press, 1965). The example of evergetism is taken from P. Veyne, *Le pain et le cirque* (Paris: Seuil, 1976), and the observation on the solidarity of Elizabeth I with Philip II from A. Somerset, *Elizabeth I* (New York: Anchor Books, 1991), p. 289. The example of Italian latrines is taken from M. Thompson, *The White War* (New York: Basic Books, 2010), p. 151; and that of vaccination from P. Fine and J. Clarkson, "Individual versus public priorities in the determination of optimal vaccination policies," *American Journal of Epidemiology* 124 (1986), 1012–20. Data on the overuse of antibiotics are found in E. Kades, "Preserving a precious resource: rationalizing the use of antibiotics," *Northwestern University Law Review* 99 (2005), 611–75. Some of the many experiments by E. Fehr and his collaborators on the maintenance or unraveling of cooperation are discussed in E. Fehr and S. Gächter, "Cooperation and punishment in public goods experiments," *American Economic Review* 90 (2000), 980–94 and, by the same authors, "Altruistic punishment in humans," *Nature* 415 (2002), 137–40. Figure 23.3 is reproduced with permission from H. Gintis *et al*., "Moral sentiments and

material interests: origins, evidence, and consequences," in H. Gintis *et al*. (eds.), *Moral Sentiments and Material Interests* (Cambridge MA: MIT Press, 2006). Evidence on the effects of closed shops on trade unions is in D. Ellwood and G. Fine, "The impact of right-to-work laws on union bargaining," *Journal of Political Economy* 95 (1987), 250–73. The punishment of do-gooders is documented in B. Herrmann, C. Thöni, and S. Gächter, "Anti-social punishment across societies," *Science* 319 (2008), 1362–7. The difference between disapproval and punishment is the topic of A. Leibbrand and R. López-Pérez, "Different carrots and different sticks: do we reward and punish differently than we approve and disapprove?" *Theory and Decision* 76 (2014), 95–118. Non-importation and non-consumption movements in America are studied in T. Green, *The Market Place of Revolution* (Oxford University Press, 2004). The open internet access to income and tax data in Norway and its consequences are discussed in J. Slemrod, T. Thoresen, and E. Bø, "Taxes on the internet: deterrence effects of public disclosure" (2013), available at: www.cesifo-group.de/ifoHome/publications/working papers/CESifoWP/CESifoWPdetails?wp_id=19075157. The snowballing effects in Eastern Europe are studied in more detail by R. Petersen in Chapter 8 of *Resistance and Rebellion* (Cambridge University Press, 2001). The quantitative threshold model on which I draw here is due to M. Granovetter, "Threshold models of collective behavior," *American Journal of Sociology* 83 (1978), 1420–43, further explored by T. Kuran, *Private Truths, Public Lies* (Cambridge MA: Harvard University Press, 1995).

# 24 Collective decision making

Often, members of a group – from the family to society as a whole – need to regulate matters of common concern by making decisions that are binding on them all. Consider again the question of regulating water consumption during periods of scarcity (Chapter 5). Sometimes, this collective action problem may be resolved by decentralized decisions, through a combination of moral, quasi-moral, and social norms.[1] Often, however, the city council has to limit the water supply or reduce consumption by banning certain uses, such as watering the lawn or filling up swimming pools. When collective action fails, collective decision making may be required.

For another example, take the practice of voting in national elections. As explained in the previous chapter, the choice whether to vote or stay home is a classic collective action problem. Knowing that his or her voice makes virtually no difference to the outcome, each citizen has a personal interest that dictates abstention. Yet if everybody abstained, or if voting dropped to very low levels, democracy itself might be in danger of being replaced by a dictatorship or an oligarchy, against (almost) everybody's interest. In many democracies voting does in fact reach respectable levels, from 50 percent to 80 percent, as a result of decentralized decisions by the citizens. Some may ask themselves, "But what if everybody abstained?" Others may say to themselves, "Since most others bother to vote, it is only fair that I should do so too." Still others may calculate that "although the impact of my vote on the viability of democracy is tiny, it is important if multiplied by the large number of other citizens it affects." In a small village, some may fear that "if I stay home, my neighbors will notice and express their disapproval."

If these motivations, singly and combined, prove too weak, voting may fall to disastrously low levels, in a process that is in part self-reinforcing ("Since few others bother to vote, why should I?"). To reverse this process, parliament may legislate to make voting obligatory and to impose a fine on non-voters, and submit the law to the approval of the voters in a referendum. When voting

[1] In Bogotá, though, the city government had an active role in *providing the information* needed to trigger the quasi-moral norms (Chapter 5).

on whether voting ought to be made obligatory, the citizens face a choice that is very different from the one they confront when contemplating whether to vote in ordinary elections with non-mandatory voting. The options are not "I vote" versus "I stay home" but "Everybody votes" versus "Everybody is free to stay home."[2] If many of those who prefer the second option in the first choice prefer the first option in the second choice, they will *decide collectively* to make voting obligatory, in a form of collective self-paternalism. Experiments suggest that people are even willing to make personal sacrifices for the sake of future generations, as long as free-rider behavior is excluded by majority voting.

Collective decision making is about making a *policy choice*. Before entering into the process of collective decision making, each member has certain *policy preferences*, which derive from his or her fundamental preferences together with *factual beliefs* and *causal beliefs* about ends-means relations. The basic goal of collective decision making is to *aggregate* individual policy preferences by one of three mechanisms to be discussed shortly.[3] The aggregation may also induce a *transformation* of individual policy preferences as the result of discussion, and it may create an incentive for individuals to *misrepresent* their policy preferences. The interaction among aggregation, transformation, and misrepresentation of preferences can make for considerable complexity.

In many of the cases I shall discuss, a smaller group of individuals make decisions that are binding on a larger group. Sometimes, they are delegated to do so, as representatives of or negotiators for the larger body. In that case, they may be constrained by the knowledge that their decisions will have to be ratified by their constituency, or that they will not be reelected if they fail to achieve satisfactory results. In other cases, the larger society has no power, short of a revolution, to influence those who make the decisions that shape their lives. Yet even here we may talk of collective decision making within the elite. After the fall of Stalin, there followed collective leadership by the Politburo. The Chilean junta that exercised power from 1973 to 1980 had a highly structured internal mode of collective decision making.

The three aggregation mechanisms I shall consider are *arguing*, *bargaining*, and *voting*. I believe that this is an exhaustive list, although in some cases the distinction may be blurred. Toward the end of the chapter I note some cases in which the distinction between arguing and bargaining breaks down.

[2] To put it differently, "I stay home but everybody else has to vote" will not be among the options in the referendum.

[3] Although the phrase "aggregation mechanism" is usually reserved for voting procedures, I use it here to denote any process in which actors who may have initially different preferences interact to bring about a decision that all of them accept as binding.

Each method can be used in collective decision making involving any number of agents. In discussing bargaining, though, I shall limit myself to two-person cases. Although multiperson bargaining certainly occurs, as in the formation of a coalition government or the allocation of emission quotas to nations, the processes are not well understood. In some three-person bargaining games, essentially any strategy combination is a game-theoretic equilibrium. In some two-person cases, *only* bargaining is effective. If two agents cannot reach agreement by arguing, voting will obviously not solve the problem, so only the bargaining procedure remains.

Let me first give some examples of the three procedures. Pure argument is observed (or at least is supposed to be the rule) in juries for which unanimity is required. Even here, some jurors may resort to tacit bargaining by virtue of their greater ability to hold out, that is, their lesser impatience to get out of jury work and back to their ordinary life.[4] Because time always matters when a decision has to be made, and because the participants in the process often discount the future at different rates, this case may in fact be typical.

Pure bargaining is illustrated by sequential "divide-a-dollar" games in which the parties make successive offers and counteroffers. The outcome is determined by the bargaining mechanism and the bargaining power of the parties, that is, the resources that enable them to make credible threats and promises. The process is illustrated in Figure 19.2.

Pure voting was Rousseau's conception of collective decision making. The citizens were to form their preferences in isolation from one another so as not to be contaminated by eloquence and demagogy. Because they would also cast their votes in isolation from one another, vote trading would be excluded. In actual political systems this ideal is never realized. It may be illustrated, perhaps, by certain low-stake decisions such as the election of members to a scientific academy whose main function is to elect new members.

The decisions to which these methods are applied vary widely. Below I give examples ranging from two parents bargaining over child custody to voting over the location of an airport. Most surprisingly, perhaps, all three procedures have been used to settle, or attempt to settle, conflicts over religious dogma. The Christian definition of the Supreme Being was reached by majority *voting* among the bishops assembled at Nicaea in AD 325. In 1561, Queen Regent Catherine of France called a colloquium at Poissy, in which leading Protestants and Catholics *argued* over the dogma of transubstantiation. At one point they seemed close to reach an agreement, but in the end they failed. The wars of religion that were then unleashed were resolved by *bargaining* between Henri

[4] In early English jury trials the practice of starving the jurors (or having them pay for their own food) until they reached a unanimous decision may also have conferred greater bargaining power on some than on others.

de Navarre (the later Henri IV) and four bishops in 1593. While accepting many Catholic tenets, Henri refused to accept the doctrine of purgatory and expressed reservations about the permanent "real presence" of Christ in the sacramental bread outside the hours of church services. At the Canterbury convocation of bishops in 1532 that preceded Henry VIII's break with Catholicism, reformers and traditionalists used the common bargaining technique of *splitting the difference*. According to Hume, "the two sects seem to have made a fair partition, by alternately sharing the several clauses."

In addition to cases involving only one of the three methods, there are many mixed cases.

Mixed arguing and voting, without bargaining, may be illustrated by hiring and tenure decisions in a university department. These are supposed to be governed only by deliberation about the merits of the candidate followed by a vote. Although this ideal does not always correspond to the reality, it sometimes does. In good departments there is a norm against logrolling, reinforced by a norm against voting without explaining one's vote.

Mixed arguing and bargaining, without voting, is illustrated by collective wage bargaining. When a union and management are deciding how to divide the income of the firm, it might appear as if only bargaining is taking place. On closer inspection, however, there is always a substantial amount of arguing about factual matters, such as the financial well-being of the firm and the productivity of the labor force.

Mixed bargaining and voting was institutionalized in the British Wages Councils and Boards in the 1950s. The possibility of a vote shaped wage bargaining even though in most cases no vote took place. The crucial factor was the presence at the bargaining table of an uneven number of independent members, along with equal numbers of members representing employers and workers. The first group served both as a mediator between the two others in the course of the bargaining process and, by virtue of their uneven number, as a guarantee that the wage would be settled by a decisive vote if no negotiated agreement was reached.

Political decision making, whether by a committee, an assembly, or the population at large, often involves all three procedures.[5] Again, this fact follows from the need to reach a decision sooner rather than later. Voting tends to arise when an issue has to be decided urgently, so that the participants do not have the time to deliberate until they reach unanimity. More prosaically, they may not be motivated to search for unanimity. If the decision is more urgent for some participants than for others, the possibility of bargaining also arises, since those who can better afford to wait may demand concessions in

[5] Even general elections may offer scope for bargaining. If voting is public, voters and candidate may haggle over the price of votes.

exchange for an early decision. In standing committees and assemblies, bargaining also arises through logrolling, which is due to unequal intensity of preferences over the issues to be traded off against each other. Other bargaining mechanisms in legislatures include filibustering, endless amendments, and "the politics of the empty chair" by which a group may exploit the rules of quorum to obtain what they could not achieve by other means.

In such cases, the sources of bargaining power are created within the assembly itself. In other cases, the decision makers can draw on resources that exist independently of the assembly – money and manpower. In 1789, the debates in the French constituent assembly were suspended between the king's troops and the crowds in Paris. In 1989, the quasi-constitutional or preconstitutional Round Table Talks in Poland were suspended between the threat of Soviet intervention and the prospect of a general strike and economic paralysis. If a vote cannot be bought with the promise of another vote, as in logrolling, it can be bought with money, for instance, with the allocation of party funds for purposes of reelection campaigning.

As will be clear from this discussion, the three modes of collective decision making may be seen as three steps in an idealized sequence, in the sense that each of them arises naturally from the preceding one. Although arguing intrinsically aims at unanimity, in the sense that it is based on reasons that are supposed to be valid for all, this end is rarely achieved. To settle the issue, voting is needed. Because voting often takes place among individuals who have many issues to decide on, it naturally gives rise to bargaining in the form of logrolling.

### Arguing

Arguing is the effort to persuade by reason giving. Ever since Pericles' eulogy of Athens, this mode of decision making has been closely linked to democratic politics:

> Our public men have, besides politics, their private affairs to attend to, and our ordinary citizens, though occupied with the pursuits of industry, are still fair judges of public matters; for, unlike any other nation, we regard the citizen who takes no part in these duties not as unambitious but as useless, and we are able to judge proposals even if we cannot originate them; instead of looking on discussion as a stumbling-block in the way of action, we think it an indispensable preliminary to any wise action at all.

The link between the institution of public debate and "wise action" can be somewhat indirect. Often, the main effect of the public setting is to exclude overt appeals to interest. In a public debate, a speaker who said, "We should do this because it is good for me" would not persuade anyone, and would, moreover, be subject to informal sanctions and ostracism that would make

her less effective in the future. Even those who are motivated solely by interest are constrained by the public setting to present their policy proposals as motivated by more impartial values. This process of *misrepresentation* of preferences differs from that of *transmutation* (Chapter 9) in the same way as deception differs from self-deception. It is the interest of the speaker, not her need for self-esteem, that causes her to misrepresent her interest as reason. Her interest may also cause her to make the misrepresentation hard to perceive, by arguing in impartial terms for a policy that deviates somewhat (but not too much) from the one that would coincide perfectly with her interest. The misrepresentation might, in fact, backfire if it were too obvious. In addition to this "imperfection constraint," speakers are subject to a "consistency constraint." Once a speaker has adopted an impartial argument on opportunistic grounds she cannot easily abandon it if, on another occasion, it no longer matches her interest. Hence *the need to disguise one's fundamental preference may induce a shift in one's policy preference*, by what we may think of as "the civilizing force of hypocrisy."[6] The seventeenth-century American minister Roger Williams put it starkly, when he wrote that Pharisees make good citizens.

To illustrate the imperfection constraint, we may first note that in many societies, property has been used as a criterion for suffrage. One may, to be sure, offer impartial arguments for this principle. At the Federal Convention, Madison argued that stringent property qualifications for the Senate, rather than protecting the privileged against the people, were a device for protecting the people against itself. But as noted, there is something inherently suspicious about such arguments, which coincide too well with the self-interest of the rich. It may then be useful to turn to literacy, as an impartial criterion that is *highly but imperfectly* correlated with property. At various stages in American history literacy has also served as a legitimizing proxy for other unavowable goals, such as the desire to keep blacks or Catholics out of politics. American immigration policy has also used literacy as a proxy for criteria that could not be stated publicly. Proposals to screen immigrants by testing them for literacy in their native language were usually justified as a way of selecting on the basis of individual merit, a widely accepted impartial procedure. The real motivation of the advocates of literacy was, however, prejudice or group interest. Patrician nativists wanted to exclude the usually illiterate immigrants from central and

[6] Might the outward expression of a hypocritical belief also induce inward endorsement? Commenting on the persecution of heretics under Henry VIII, Hume wrote that the practice "may, indeed, seem better calculated to make hypocrites than converts; but experience teaches us, that the habits of hypocrisy often turn into reality; and the children at least, ignorant of the dissimulation of their parents, may happily be educated in more orthodox tenets." I am more persuaded by the latter part of the argument than by the former, since persecution offers the victim a *reason* to dissimulate (Chapter 9).

southeastern Europe. Labor feared that an influx of unskilled workers might drive wages down.

To illustrate the consistency constraint, I shall cite some arguments used in wage bargaining. As I note later in the chapter, the outcome of wage bargaining is often shaped by the raw material bargaining power of the parties. It can also, however, be affected by norms of fairness. If one party adopts a norm on opportunistic grounds, it may, however, be stuck with it. Thus if a union successfully argues for a wage rise by citing a norm that windfall gains for the firm should be shared with the workers, it may find it difficult to resist the argument that windfall losses should also be shared. Conversely, once the Great Depression ended, many firms regretted the ability-to-pay argument they had used to keep wages down. When, in the 1930s, the wages of Swedish metal workers lagged behind those in the construction industry, they appealed successfully to a solidaristic wage policy to reduce the differential. Later, when the metal workers became the high-wage outliers, they were bound by their past appeals to solidarity.

It would obviously be wrong to think that arguing can *always* be reduced to more or less subtle ways of promoting one's interest. If that were the case, there would be no point in misrepresentation since nobody would ever be taken in. If speakers are motivated by a sincere desire to promote the public good, argument and debate may change their beliefs in ways that induce a change in policy preferences. This is especially likely to occur if the various members of a group have access to different information, so that they can improve the quality of their decisions by pooling their knowledge.[7] If the body is a representative one, it is then important to select delegates with widely different backgrounds. In electing representatives to a national assembly, for instance, this consideration speaks in favor of proportional voting with a low threshold or no threshold at all.[8] One might also require representatives from an electoral district to be residents of that community.

People also, although perhaps more rarely, argue about fundamental preferences and change them as a result of debate. Often, change occurs through the discovery of hidden similarities between cases or the exposure of superficial similarities. Many people are opposed, for instance, to the mandatory use of "cadaver organs" for transplantation purposes. They believe that if the family has religious objections to this procedure, their feelings ought to be respected. Against this view one might point to the mandatory use of autopsies in the case

[7] Recall, however, that this improvement is more likely if they pool their raw data than if they simply pool the conclusions reached on the basis of raw data (Chapter 22).

[8] Other considerations, notably the need for effective governance, may speak in favor of majority voting or proportional voting with a high threshold. In elections to constituent assemblies, in which governance is a secondary consideration, there is a tendency to choose delegates by proportional voting.

of suspicious deaths, even when the procedure is contrary to the religious beliefs of the family. If invasive measures are in order to determine the cause of death, one might argue that they should also be acceptable for the purpose of saving lives. Change can also occur when a general principle is seen to contradict intuitions about particular cases. A person might accept the mandatory use of cadaver organs on utilitarian grounds but balk at the implication that one would be justified in killing a randomly chosen person and using their heart, kidneys, lungs, and liver to save the lives of five others.[9] As a result, an initial unqualified utilitarianism might be revised to take account of non-consequentialist values (Chapter 4).

The benefits of arguing may be undermined, however, by the effects of *speaking before an audience*. Public-minded individuals may, no less than others, be subject to amour-propre that makes them reluctant to admit in public that they have changed their mind. In Chapter 3 I noted that this was the main reason Madison gave, long afterward, why the Federal Convention was held behind closed doors with silence imposed on the delegates. His argument might, however, seem to conflict with a traditional argument for opening assembly debates to the public. Many legislative decisions have a strong short-term impact on legislator interests. If the decision-making process is shielded from the public eye, arguing about the common interest will easily degenerate into naked interest bargaining. Allowing the public to follow the proceedings and observe the votes tends to limit such self-serving scheming and, as a by-product, promote the public good. As Bentham wrote, "The greater the number of temptations to which the exercise of political power is exposed, the more necessary is it to give to those who possess it, the most powerful reasons to resist it. But there is no reason more constant and more universal than the superintendence of the public." Or as the American judge Louis Brandeis said, "Sunlight is the best disinfectant."

These remarks point to a tension in the process of arguing. If debates are held in public, the quality of argument will suffer. If they take place behind closed doors, arguing may degenerate into bargaining. The tension may be attenuated, however, if the matters to be decided leave little room for the play of private interest. Constituent assemblies may be less prone to self-serving decisions than ordinary legislatures, not because the delegates are more impartially motivated but because (or to the extent that) their interests have less purchase on the issues at hand.

[9] Utilitarians tend to deny that this implication follows. They argue, typically, that the negative effects of the fear and uncertainty that would be generated by the knowledge that one might be chosen as a "random donor" would more than offset the benefits of the practice. *But how do they know this?* I suspect that they reason backward, from the obvious unacceptability of the practice to the existence of costs that would exclude it on utilitarian grounds, rather than forward, from the demonstration of costs to the rejection of the procedure.

## Voting

Voting may be needed when arguing fails to generate a consensus on policy. Voting systems vary greatly. In popular voting, dimensions of variation include suffrage, eligibility, the mode of voting (secret versus open), the majority needed for a decision, and, in most referendum systems, the quorum. In assembly votes, the main dimensions are the quorum, the size of the majority, and the choice between roll-call voting and a show of hands (and similar procedures, such as "shouting" or "sitting and standing"). Secret voting in assemblies is rare, but not unheard of. It occurred in the French parliament between 1798 and 1843, and in Italy between 1948 and 1988. Most assemblies choose their presidents by secret ballot. Note that secret voting is to be distinguished from closed proceedings in which no visitors are allowed. The latter may be combined with public voting that enables the assembly members to make credible promises of logrolling, which would be impossible with secret voting. By contrast, if the proceedings are open to the public some auditors may have a negative reaction if they see their representatives voting against their preference on one issue, since they cannot observe the gains thus made possible on another.

In the following I restrict myself to majority voting. Even though this is not a universal practice, the decision to adopt proposals by a larger majority such as three-fifths or two-thirds would itself, it seems, have to be made by simple majority. Constituent assemblies, which often impose qualified majorities for future constitutional amendments, almost invariably use simple majority voting in their own proceedings.[10] The idealized model in which constituent assemblies behind the veil of ignorance decide by unanimity that they will decide by majority voting once the veil is lifted has little relevance for actual constitution making. I shall ignore the issue of quorum, except to note that abstention or the "politics of the empty chair" can be used by a minority to block a decision that would have passed had it shown up and voted against it.

Voters may differ in their beliefs as well as in their ultimate goals (see the example of the French debate on bicameralism in 1789 that I discuss later). In other cases, they may be similar in one of these two respects and differ in the other. Since what is actually observed and aggregated are policy preferences, it may be hard to disentangle the two factors that go into their making. In the

[10] The making of the South African constitution of 1996 is a partial exception. The requirement that it be adopted by a qualified majority was laid down in the interim constitution of 1993, which was itself adopted by bargaining rather than voting. Another exception is the making of the Norwegian constitution of 1814, when the assembly decided that any proposal that garnered more than two-thirds of the votes would be definitely adopted, that is, not be subject to revisions in later sessions. Thus the framers deviated from a principle that is normally followed in constituent assemblies: since the effect of a given clause in the constitution often depends on the other clauses, *nothing is settled until everything is settled*.

abstract, we may nevertheless try to determine the effects of majority voting on the aggregation of beliefs (assuming identical goals) and on the aggregation of fundamental preferences (assuming identical beliefs). According to Tocqueville, democracy (i.e. majority voting with a large franchise) was the best system for determining the *ends* of public policy, but a poor system for determining the *means* to those ends. Democratic officials may "commit grave errors" but "will never systematically adopt a line hostile to the majority."[11] Earlier, James Harrington had observed that the people would not "cast themselves into the sea" as a mad prince might do.

Consider first aggregation of beliefs. There is a long-standing debate whether an extended or a narrow franchise is the better procedure for arriving at correct beliefs – whether the many are wiser than the few. According to Aristotle, this was a matter of weighing quantity (the number of participants in the political process) against quality (the competence of the participants):

> Quality may exist in one of the classes which make up the state and quantity in the other. For example the low-born may be more numerous than the noble or the poor more than the rich, yet the more numerous class may not exceed in quantity as much as they fall behind in quality. Hence these two factors have to be judged in comparison with one another. Where therefore the multitude of the poor exceeds in the proportion stated [so as to offset their inferior quality], it is natural for there to be a democracy.

In modern language, the issue can be stated in terms of Condorcet's "jury theorem." Suppose that members of a jury state their beliefs about whether the accused has in fact done what the prosecutor claims he did, and that each of them has a greater than 50 percent chance of being right. Condorcet showed that if the jury decides by majority voting and the members form their opinions independently of each other, its chance of getting it right increases with the size of the jury,[12] and converges to certainty when the jury becomes indefinitely large. Also, for a given size of the jury, the chance of the majority's getting it right increases when the chance of each jury member's getting it right goes up.[13] Hence, as Aristotle suggested, one may improve the outcome either by increasing the number of jurors or by increasing their qualifications.[14]

[11] He does not ask, though, whether the *occasional* liability to *grave* mistakes might not be more serious than the *systematic* bias of a non-democratic regime. Although Tocqueville claimed that because of its favorable geographical situation the United States could afford to make mistakes, that might not be true of other countries.

[12] Assuming, contrary to what is argued in the next paragraph, that the likelihood of each voter's being right is unaffected by an increase in the number of voters.

[13] The chance of the majority's getting it right also increases if one requires a qualified majority, such as 60 percent. In that case, however, one might get a "hung jury" in which neither the guilt nor the innocence of the accused gathers the required majority.

[14] One might also try to ensure that the conditions of Condorcet's theorem obtain by making it more likely that the beliefs of the voters are in fact independent of each other. In this perspective, Rousseau's proposal to ban discussion prior to deliberation might make sense.

Going beyond Aristotle, we can observe that qualifications may be a direct function of number rather than of socioeconomic position. In social-science language, the competence of voters may be "endogenous" to the system rather than given "exogenously." Suppose one has to choose between two political systems, oligarchy and democracy, both deciding by majority vote but with different size of the franchise. In a democracy voters will rationally decide to remain ignorant, since the impact of each on the outcome is very small.[15] In an oligarchy, voters will invest more in gathering information since each of them has a larger impact.

Bentham noted that this argument also applies to voting in an assembly: "The greater the number of voters the less the weight and the value of each vote, the less its price in the eyes of the voter, and the less of an incentive he has in assuring that it conforms to the true end and even in casting it at all." In responding to the argument that an assembly (he had in mind the French Assemblée Constituante of 1789) ought to be numerous, since "the probability of wisdom increases with the number of members," he wrote that "the reduction that this same cause brings in the strength of the motivation to exercise one's enlightenment offsets this advantage." In this quality–quantity trade-off there will be an optimal size of the electorate or the assembly that maximizes the probability that majority voting will yield the correct belief.[16] Whether that optimum can be effectively determined, is another matter.

Consider next aggregation of preferences by voting. People may have an incentive to vote for other proposals or candidates than those they would most prefer to see adopted or elected. The choice of the open rather than the secret ballot can induce this phenomenon. In classical Athens, most decisions by the assembly were by a show of hands, with the result that some citizens may have been afraid to vote their minds. Thus Thucydides states that "with the enthusiasm of the majority [for the Sicilian expedition], the few that liked it not, feared to appear unpatriotic by holding up their hands against it." The choice of roll-call vote rather than other methods, such as "standing versus sitting," can also intimidate voters. In the constituent assemblies in Paris (1789–91) and

At the same time, if deliberation improves the quality of beliefs, it cannot be an objection that it also makes them less independent of each other. The conditions of the theorem are sufficient for majority voting to produce a good outcome, but not necessary.

[15] Since the decision to vote may itself be irrational (Chapter 14), one might ask whether citizens might not also irrationally invest in information about the issues at stake. In the present context, however, the relevant issue is whether citizens invest more when the franchise is narrow than when it is wide, just as more voters may turn out when the election is seen as close.

[16] In the abstract the optimum could be at one of the extreme ends – either a single individual or all adult persons. Under reasonable assumptions, there is more likely to be an "interior maximum." If the optimal size is small, one might choose the voters at random among the citizens at large to ensure that they do not represent sectarian interests. In this perspective, voting would be a *function* rather than a *right*.

in Frankfurt (1848), radicals routinely demanded roll-call votes in important matters, with the implicit and sometimes explicit threat that they would expose those who voted against radical proposals to popular violence by circulating lists of their names. Even if there was a clear majority under the "standing versus sitting" system, which made it difficult to identify how individuals cast their vote, the outcome might be reversed upon roll-call vote.

Misrepresentation can also arise with the secret ballot. In essentially all voting systems situations can arise in which a voter, by voting for an alternative other than her first-ranked one, can bring about an outcome that she prefers to the one that would have occurred had she voted sincerely. (An exception could arise if candidates or proposals were chosen by a randomizing device, with the probability of an alternative's being chosen equal to the proportion of voters favoring it. In this case, the problem of the "wasted vote" would not arise. The disadvantages of the system are obvious and explain why it has never been chosen.) The desire to see one's first-ranked option win by a margin that is not too wide may induce one to vote against it. In Chapter 17 I mentioned, for instance, how Socialists might vote for Communists in order to move the platform of their party to the left. If it is certain that one's first-ranked option will not be chosen, one may vote for the best alternative of those that have some chance of winning. Some voting systems also create an incentive to rank a candidate or proposal preferred by other voters less favorably than one's real preferences would dictate (see the example later), or to introduce new alternatives for the sole purpose of making the choice of one's preferred alternative more likely.

### How voting differs from individual decisions

An individual decision is based on the desires and beliefs of the agent. I have assumed that she *knows* what she wants and what she believes, in other words that the premises for her decision are *determinate*. The decision itself may of course be indeterminate, not in the sense that she does not make any, but in the sense that it is not determined uniquely by the premises. In the presence of uncertainty, for instance, the agent may not know what to do and decides by flipping a coin. Collective decisions based on majority voting can be indeterminate in the more fundamental sense that, metaphorically speaking, the group *does not know what it wants* or *does not know what it believes*. These expressions are metaphors, since only individuals can have wants and beliefs. Yet it might seem natural and harmless to impute wants and beliefs to a group by determining the majority preference and the majority belief. It has been known since 1785 and 1837, however, that the notions of a majority preference and of a majority belief can be

Table 24.1

| | *Businesspeople* | *Workers* | *Professionals* |
|---|---|---|---|
| Golf course | 1 | 2 | 3 |
| Orchestra | 2 | 3 | 1 |
| Pool | 3 | 1 | 2 |

indeterminate. I shall name these situations, after their discoverers, the *Condorcet paradox* and the *Poisson paradox*.[17]

The *Condorcet paradox* arises when the outcome of majority voting is indeterminate. Suppose there are three blocs of roughly equal size in a municipal assembly, representing, respectively, the business community, industrial workers, and social service professionals. The assembly is to choose among building an indoor swimming pool, subsidizing the local symphony orchestra, or building a golf course. Conforming to the stereotype of these groups, suppose that (after long debates) they rank the options as shown in Table 24.1.

If the alternatives are held up against each other in pairwise votes, there is a majority of businesspeople and workers who prefer the golf course to the orchestra, a majority of businesspeople and professionals who prefer the orchestra to the pool, and a majority of professionals and workers who prefer the pool to the golf course. Hence the "social preferences" are *intransitive* or *cycling*. In the case of individual choice, transitivity was a requirement of rationality (Chapter 13). In the present context, the question is not so much one of rationality as of determinacy. If all the municipal council has to go by is the rankings in Table 24.1, it is hard to see how they could make any decision at all. Since the vote was taken because the council was unable to reach consensus, more debate is unlikely to help. If one could measure the *intensity* with which the various groups prefer one of the options to another, or the extent to which the options satisfy objective needs, one might be able to say that one option was unambiguously superior to the others. There is no general procedure, however, that allows us to compare degrees of preference intensity or of need satisfaction across individuals. Asking them how much they value the options, for instance, by making them rank them on a scale from 0 to 10, is pointless. We cannot know whether a given score (e.g. 7) means the same thing for members of the three groups. Also, asking them to rank the options would give them an incentive to misrepresent the intensity of their preferences,

[17] What I call the Poisson paradox is more usually named the "doctrinal paradox" or the "discursive dilemma," terms coined by legal scholars and philosophers who rediscovered it in the 1980s and 1990s. Earlier, it was rediscovered in 1921 by the Italian legal scholar Vacca.

for example by assigning 10 to their top-ranked option and 0 to the main rival even if it is in fact their second choice.

It is not clear how important this problem of "cycling social preferences" is in practice. It cannot arise if individual preferences are "single peaked," meaning that the options can be ranked from "highest" to "lowest" in such a way that the preferences of each individual are steadily increasing toward his or her most preferred policy and steadily decreasing as one moves away from it. In many cases, this is a reasonable property of preferences. If an individual's preferred tax schedule is 20 percent, he or she will prefer 19 percent to 18 percent and 21 percent to 22 percent. Moreover, there are no instances of an assembly's simply throwing up its hands and declaring that because there is no "popular will" no decision will be made. In fact, if the status quo is one of the options, this idea is incoherent. Some decision is always made, whether by default (retaining the status quo), by adoption of a traditional voting procedure, or by manipulation of the agenda.

Yet the fact that a decision is reached does not imply that it embodies the popular or "general" will in some non-arbitrary sense. For a constellation of (sincere) preferences such as the one given in Table 24.1, the very idea of a general will is meaningless. How often do such constellations occur? Political scientists have offered a number of examples. Others have argued that the alleged examples have been misdescribed, and that a closer examination refutes these specific claims about cycling majorities. I shall describe two cases that appear to be genuine instances of cycling preferences.

On October 8, 1992, the Norwegian parliament decided that the future airport for the Oslo area should be located at Gardermoen (I shall refer to this option as alternative G). Other options were Hobøl (alternative H) and a solution that involved a combination of Gardermoen and the existing Fornebu airport (alternative D). The options were not to be held up against each other but considered successively against the status quo. Once an option received a majority of the votes, it was adopted. Although this serial voting was the traditional voting system in the parliament, other systems are possible, for example, holding the options up against each other in pairwise votes until one winner remains. With successive voting, the order in which the options are voted on can be decisive, as we shall see shortly.

The *expressed* party preferences, which with unimportant exceptions coincided with the votes of the deputies, were as follows:

The Labor Party (63 deputies): $G > D > H$
A coalition of the Socialist Left Party, the Christian Democrats, and the Agrarian Party (42 deputies): $D > H > G$
The Conservative Party (37 deputies): $H > G > D$
The Progress Party (22 deputies): $H > D > G$
One independent deputy: $G > H > D$

Assuming these to be the *sincere* preferences, social preferences were cycling: D beats H 105 to 60, H beats G 101 to 64, and G beats D 101 to 64. Before voting, parliament voted on the order in which the alternatives should be considered. Labor proposed G-D-H, whereas the president of the parliament proposed D–H–G. When the proposals were held up against each other, Labor's won. If the president's proposal had won, Labor would probably had voted for D, since otherwise its failure to garner a majority for D would have led to the adoption of its bottom-ranked proposal, H. Under the order that was adopted, the Conservative Party was in a similar predicament. In the end, the Conservatives voted for G, since if they had voted against it, their bottom-ranked proposal, D, would have won. Although it is abstractly possible that Labor was insincere in stating D as its second-ranked option, and that it did so only to make the Conservative Party believe that voting against G would lead to the adoption of D, there is no evidence to that effect. If this was in fact the case, social preferences would not be cycling, since H would beat both D and G.

In the second example it is pretty much excluded that the cycling preferences could be a mere artifact of misrepresentation. It arose in the context of deciding the order of demobilization from the American army after World War II. Getting out early was a scarce good, which had to be allocated fairly. To determine the criteria, the army conducted large-scale surveys among the enlisted men. In a survey in which the criteria were held up against each other in pairwise comparisons, the rankings showed some collective inconsistency. Thus 55 percent thought that a married man with two children who had not seen combat should be released before a single man with two campaigns of combat; 52 percent rated eighteen months overseas as more important than two children; and 60 percent rated two campaigns as worth more than eighteen months overseas. It is most unlikely that the respondents were misrepresenting their preferences.[18]

*The Poisson paradox.* In a book on the statistical analysis of legal decisions, the French mathematician Poisson inserted, perhaps as a curiosum, this footnote:

> Two individuals, whom I shall call Pierre and Paul, are accused of theft; to the question whether Pierre is guilty, four jurors say yes, three others yes, and the five remaining no: the defendant is declared guilty by a majority of seven votes to five; to the question

[18] The authors of the study from which I take these findings wrote that "a high degree of internal consistency on such intricate hypothetical choices was hardly to be expected," suggesting that the problem was one of individually inconsistent rankings. If the majorities had added up to more than 200 percent, this suggestion would have been justified. As they add up only to 167 percent, it is quite possible that the rankings were individually consistent and yet gave rise to a collective intransitivity. The study was published in 1949, two years before Kenneth Arrow's pathbreaking work on preference aggregation and the inconsistencies to which it is vulnerable.

Table 24.2

| | *Fundamental Preferences* | *Beliefs* | *Policy preferences* |
|---|---|---|---|
| Reactionaries | Destabilize the regime | Bicameralism will stabilize the regime | Unicameralism |
| Moderates | Stabilize the regime | Bicameralism will stabilize the regime | Bicameralism |
| Radicals | Stabilize the regime | Bicameralism will destabilize the regime | Unicameralism |

whether Paul is guilty, the first four jurors say yes, the three others who had said yes against Pierre say no against Paul, and the five remaining say yes: Pierre is therefore declared guilty by a majority of nine votes to three. Next one asks whether the theft has been committed by several individuals, which in case of an affirmative answer entails a more serious punishment. Following their previous votes, the first four jurors say yes and the remaining eight who had declared either Paul or Pierre to be innocent, say no. Hence even though there is no contradiction in the votes of the jurors, the decision of the jury is that both are guilty of theft and that the theft has not been committed by several individuals.

The jurors could reach their decision on the issue of joint guilt by two procedures. If they voted directly on this issue, a majority would find Not Guilty. If they voted on the question of the guilt of each individual and then drew the logical conclusion from these two votes, a majority would find Guilty. These are usually referred to as the "conclusion-based" and the "premise-based" procedures. Since both seem equally plausible, we might say that the jury does not know what it believes. Although Poisson may have believed the paradox to be a mere curiosum, it has been shown to occur quite frequently, notably in the deliberations of multi-judge courts.

A similar paradox can arise when a group has to aggregate both beliefs and preferences. As an example, consider the debates over unicameralism versus bicameralism in the French Assemblée Constituante of 1789. Very broadly speaking, the assembly contained three roughly equal-sized groups. The reactionary right wanted to set the clock back to absolute monarchy, the moderate center wanted a constitutional monarchy with strong checks on parliament, and the left wanted a constitutional monarchy with weak checks on parliament. On the issue of bicameralism, the constellations were, highly simplified, as shown in Table 24.2.

In the end, bicameralism was defeated by the alliance of reactionaries and radicals. This general phenomenon – policy agreement based on preference differences and belief differences that cancel each other – is quite common. One might even achieve unanimity on that basis,[19] although obviously of a

[19] In the French assembly, this outcome occurred in May 1791 when radicals, moderates, and reactionaries joined forces in voting for a law that made the members of the constituent assembly ineligible for the first ordinary legislature. The aim of the radicals was to weaken

different kind from the one that might emerge in the "ideal speech situation" in which speakers are motivated only by the common good and are willing to listen to argument.

In my stylized rendering of the debate, a majority *believed* that bicameralism would stabilize the regime and a (different) majority *wanted* to stabilize the regime (see Table 24.1). If collective decisions had been made by first aggregating beliefs by (sincere) majority voting, next aggregating fundamental preferences by (sincere) majority voting, and finally taking the action that according to the aggregate belief would best realize the aggregate preference, *bicameralism* would have been the choice.[20] The actual decision was taken by voting directly on the conclusion, and *unicameralism* was adopted.

To my knowledge, assemblies always vote directly on proposals, never on premises or "reasons" for the proposals.[21] As I mentioned, smaller groups can use either procedure. For a realistic instance in which the Poisson paradox might arise in a mixed belief–preference aggregation, consider a Central Bank Committee that is to decide on changes (or no change) in the key interest rate. Each member has factual beliefs about the state of the economy and normative views about the trade-off between inflation and unemployment. The final decision might depend on whether the committee uses the conclusion-based or the premise-based procedure.

### Bargaining

Bargaining occurs in a situation of mixed cooperation and conflict. Two parties are in a situation where they can make each other better off by cooperating. There are, however, many such mutually improving arrangements, with unequal benefits to the two parties. Each party will therefore try to obtain a cooperative arrangement that is favorable to himself. The basic dilemma of bargaining is that many tactics and strategies that bargainers use to obtain an agreement that are favorable to themselves tend to delay the agreement and to impose other costs of bargaining that reduce the size of the pie that is to be divided. In the words of one scholar, "Bargaining has an inherent tendency to eliminate the potential gain which is the object of the bargaining." An

the legislature in favor of the club of the Jacobins; that of the reactionaries to weaken it in favor of the king. The vote was unanimous, since the moderate center, "drunk with disinterestedness" (Chapter 5), enthusiastically voted to deny themselves a role in the future legislature.

[20] To prevent this outcome, the reactionaries could have falsely stated a belief that bicameralism would destabilize the regime, thus creating a majority for that belief and hence a majority for the choice of unicameralism.

[21] On December 2, 1882, the House of Commons adopted a resolution requiring that any vote on the reasons for a piece of legislation be taken after the vote on the law itself. According to Robert's *Rules of Order*, "It is usually inadvisable to include reasons for a motion's adoption within the motion itself."

important example is the tendency of some firms to build up large inventories for the sole purpose of being able to weather a strike.

*Threats and promises* are the main tools of bargaining. A spouse may threaten to litigate for sole custody of a child unless the other spouse agrees to joint custody. In wage bargaining, workers can threaten to strike, to work to rule, or to refuse overtime work, while employers can threaten with lockouts or plant closures. The management of a firm may threaten to fire an employee unless he works harder at his job. One country may threaten to invade another unless it makes territorial concessions. In a constituent assembly, a delegate from one territorial unit may threaten to walk out unless the assembly adopts a mode of representation that is to the advantage of that unit. American senators may threaten with filibustering to make the president withdraw a nomination. Congress may threaten to refuse to vote the budget if the president uses a veto to override legislation.

Turning to promises, a member of a group that decides by voting may promise to vote for a proposal that is important for one of her colleagues, on the condition that the latter votes for one that matters to her (logrolling). The seller of a house may promise not to begin renegotiating if a buyer meets his asking price. Similarly, a kidnapper may promise to release the victim once the ransom has been paid, rather than retaining the victim and making a new demand. Conversely, a government may promise to let a terrorist out of jail once his co-terrorists have released the victim they have kidnapped. A victim of kidnapping may promise not to describe the appearance of the kidnappers to the police if they release him. A person in a Prisoner's Dilemma situation may promise to cooperate if the other does so as well.

To be effective, that is, to change behavior, threats and promises have to be *credible*. The person who is the target of a threat or a promise has to believe that the *threat will be carried out* if he does not comply or that the *promise will be kept* if he does. In the simplest case, this belief is based on the fact that it will be in the interest of the person making the threat to carry it out or in the interest of the person making the promise to keep it. If I surprise an unarmed burglar in my house and threaten to call the police unless he leaves, he will comply because he knows it will be in my interest to do so if he does not. (If he is armed, the threat may not be credible.) In a classical example proposed by Thomas Schelling, a promise to my kidnapper not to reveal his identity if he releases me is credible if I provide him with damaging and verifiable information about myself that he would have an incentive to divulge if arrested.

A person can also make it a rule always to keep a promise or carry out a threat, even when on a given occasion it is not in her interest to do so. By this means she can build a *reputation* that will be useful over the long run. *Irrationality* can also be a boon in bargaining (but only if perceived by others). In a given situation, the threat to walk away from the bargaining table might be

credible if made in anger, but not otherwise. *Incompetence*, too, can be helpful, if a bargainer is (perceived to be) unable to see where her interest lies. Agents may also *invest in credibility,* as when President Kennedy asserted, after the fiasco of the Bay of Pigs, that "We have a problem in trying to make our power look credible, and Vietnam looks like the place." Most explanations of the Vietnam War refer to the belief by successive American administrations in the domino theory – the threat of the United States to intervene against Communist forces in Laos and other countries would not be credible if it abandoned South Vietnam. Kennedy's remark adds a twist to this explanation, by suggesting that the Vietnam War was initiated to *create* credibility and not only pursued to *maintain* it.

Let me give some examples of non-credible promises and threats. Beginning with promises, consider a failed attempt at logrolling in the French constituent assembly in the fall of 1789. In three meetings between the leader of the moderates, Mounier, and the radicals Barnave, Duport, and Alexandre Lameth, the latter three made the following proposal. They would offer Mounier both an absolute veto for the king and bicameralism, if he in return would accept that the king gave up his right to dissolve the assembly, that the upper chamber would have a suspensive veto only, and that there would be periodic conventions for the revision of the constitution. Mounier refused outright, arguably because he did not believe in the ability of the three to *deliver* on their promise, since the assembly did not have parties in the modern sense of disciplined groupings that can be made to vote as a single bloc.

For another instance, consider promises of immunity to prosecution for outgoing leaders in transitions to democracy. Promises to this effect were made, accepted, and broken in Argentina in 1983, in Uruguay in 1984, and in Poland and Hungary in 1989. (In the Latin American countries, threats of a military coup then forced compliance.) In retrospect, the generals and party leaders should have understood that these promises were not credible, since the negotiating incoming leaders could not guarantee that courts and legislatures would respect them. In Poland, the negotiators for the opposition in the Round Table Talks, who belonged to the left wing of Solidarity, argued that *pacta sunt servanda* – promises are to be kept. When the right wing of the movement gained power, they ignored the pledge. In the demobilization of the Colombia paramilitaries that began in 2003, the government's negotiators made several promises that were subsequently struck down by the courts.

In Chapter 25, I discuss why the lack of a hard-to-amend written constitution made it impossible for the British parliament in the eighteenth century to make credible promises to the American colonies. It has also been argued that, prior to the Glorious Revolution of 1688, English monarchs were hampered by their inability to make credible promises to honor their commitments to creditors. As a result, they had to pay higher interest on their loans to

compensate for the risk of default. When Robert Walpole established the Sinking Fund in 1717, "he announced that the appropriation of duties to the Fund would constitute a kind of 'fundamental law', to be considered unbreakable by future Chancellors of the Exchequer." Since no government can bind a future government, his commitment was empty. Walpole himself was the first Chancellor to violate this contract, when he transferred surpluses on the Fund to his budgetary account in 1733. French kings, too, were "impotent because omnipotent." Machault, the ablest minister of Louis XV, tried to establish a sinking fund to be used only for the payment of debts, but since he was unable to prevent the fund from being raided in times of urgency, creditors were not impressed.

In everyday life, instances of non-credible *threats* are commonplace. As any parent knows, children often call the bluff of angry parents when they announce drastic punishments for the performance or non-performance of some action. More generally, if threats are made in the heat of passion and the target knows that passions tend to decay quickly, he may dismiss them. Until the resignation of Spiro Agnew, the threat to impeach Richard Nixon was not credible, because the consequences would be unacceptably bad. In 1986, Ronald Reagan's threat to create a missile-defense system was intrinsically non-credible, except to his interlocutor in Reykjavik, Mikhail Gorbachev. The historian of their encounter, who was present at the talks, writes that "Reagan wanted so badly to build it and Gorbachev wanted so badly to stop it, that it assumed for them, and practically only for them, a reality it actually lacked." Their situation was a perfect illustration of a phrase I have quoted repeatedly, that one easily believes what one fears and what one hopes. The question whether the threats of the United States and the Soviet Union to use nuclear weapons against an attack were credible poses intriguing philosophical issues, but does not lend itself to empirical resolution. President Clinton and President Obama called the bluff of the leaders of the Republican Party when they threatened to shut down government unless certain demands were met. The North Korean threats against the universe at large would be non-credible were it not so clear they are made for internal consumption only.

I now consider two examples in more detail: wage bargaining between a firm and a trade union[22] and bargaining between two parents over child custody.

Important determinants of a negotiated wage agreement are the *outside and inside options* of the parties. In wage bargaining, a worker will not accept an offer for a lower wage than he can get at the firm across the street. This is the worker's outside option. The wage he can get by moving to another *province*

[22] Here I first consider Western-style wage bargaining. Toward the end of the chapter, I discuss the emerging wage bargaining system in China.

does not provide a lower bound on the employer's offer, since the move is costly. The wage paid to workers in another *industry* does not constitute a lower bound either, if the worker would be unqualified for a job in that industry. It may nevertheless influence the bargaining outcome through social norms. The firm, too, has outside options, such as closing down its operations and selling the plant at scrap value.[23] These outside options represent the value of the alternatives open to the parties after a *definitive* break-up of the relationship.

The firm and the worker also have *inside options*. These are the resources that enable the parties to hold out during a *temporary* break-up of the relationship, caused by a strike or a lock-out. Tocqueville noted that in France around 1830, "nearly all workers have some secure resources [a plot of land] that allow them to withhold their services when others are unwilling to grant them what they consider a just reward for their labor." In contemporary societies, the most important inside option for the workers is the strike fund. For the employers, it is the size of the inventory. To break the British coal-mining unions, Margaret Thatcher encouraged the coal-mining employers to build up a year's worth of coal inventories. The inside options of the workers improve if most of them are young men or women without families or do not have heavy mortgages on their homes. The inside option of the firm is improved if it employs labor-intensive rather than capital-intensive technology, since the latter usually requires higher interest payments on loans.[24]

Bargaining outcomes are also affected by the "formal preferences" (Chapter 5) of the agents – time discounting and risk attitudes. Generally speaking, impatient and risk-averse agents are at disadvantage. In this case, impatience is not due to pure time discounting, but to scarcity (see Chapter 6), such as lacking a small plot or a strike fund to support the workers during a strike. Since an impatient agent is willing to give up a share of the pie in order to get it earlier, she gets a smaller share than a more patient agent. Risk aversion makes workers less willing to substitute higher wages for a greater unemployment risk. The outcomes can also be affected by social or moral norms. A firm that increases its profits but does not offer a wage increase may be seen as acting unfairly, generating strong emotions of anger. Under the influence of this emotion, the workers may carry out a strike threat even though their inside options, by assumption, have not improved. Comparisons with wages in other firms are also important in shaping perceptions of fairness, even when a job in

[23] The threat of shutting down a plant permanently might seem non-credible, Yet that is what Roger Milliken did in 1956, when workers at his Darlington factory organized to form a union. I do not know whether he had threatened to do so.

[24] It is not true, therefore, as Marx said, that machinery "is the most powerful weapon for suppressing strikes." The decline in the US steel industry has been explained by the fear of investing in "hostage capital."

these firms does not constitute an outside option. If a firm pays its workers a wage of C, an increase in the legal minimal wage from A to B can induce an increase in the wage from C to D even when $C > B$.

When couples with a child or several children split up, they may not be able to agree on child custody. Before going to the courts, they may engage in private bargaining. I shall assume that there are two children, a boy and a girl, and that the parents rank the custody allocations as follows:

FATHER: Custody of both children preferred to custody of boy only preferred to custody of girl only preferred to custody of neither. We indicate the cardinal utilities of these options as $u_1$, $u_2$, $u_3$ and $u_4$.

MOTHER: Custody of both children preferred to custody of girl only preferred to custody of boy only preferred to custody of neither. We indicate the cardinal utilities of these options as $v_1$, $v_2$, $v_3$ and $v_4$.

If the parents cannot reach a negotiated agreement, they go to court. The expected outcome of the legal decision represents their outside option, or, as it is also called, the *threat point*. Their private bargaining takes place "in the shadow of the law," in the sense that neither parent will accept a custody arrangement whose utility for him or her is less than the expected utility of the court-imposed solution. Suppose for specificity that they believe it is equally likely that the court will award full custody of both children either to the one or to the other.[25] The expected utility of this solution $(u_1 + u_4)/2$ for the father and $(v_1 + v_4)/2$ for the mother. (Recall from Chapter 12 that cardinal utilities are linear in probability.) Suppose, moreover, that utilities for the possible arrangements are as shown in Figure 24.1.

In theories of bargaining, the lines connecting the four vertices represent the utilities to the parents of various custody *probabilities*. For instance, the midpoint on the line between "father gets both" and "father gets boy, mother gets girl" indicates the expected utility to the parents if they flipped a coin between these two options. Needless to say, this will never happen in actual bargaining between the parents; it is merely a device for making the situation mathematically tractable. In a more realistic interpretation, we can assume that the parents are bargaining simultaneously over child custody and the division of their financial assets. The midpoint on the line could then indicate a situation in which the mother made a financial *side payment* to the father to get custody of her daughter. None of the "pure" solutions represented by the vertices will be acceptable to both parents, since one of them will always prefer going to court. The only mutually acceptable outcomes are "mixed solutions" on the

[25] This might come about, for instance, in the unlikely event that the law says that the court should flip a coin between these two solutions, or if both parents believe that both are equally fit for custody. In many actual cases, the sum of the probability the father assigns to getting full custody and the probability the mother assigns to getting full custody probably exceeds 100 percent.

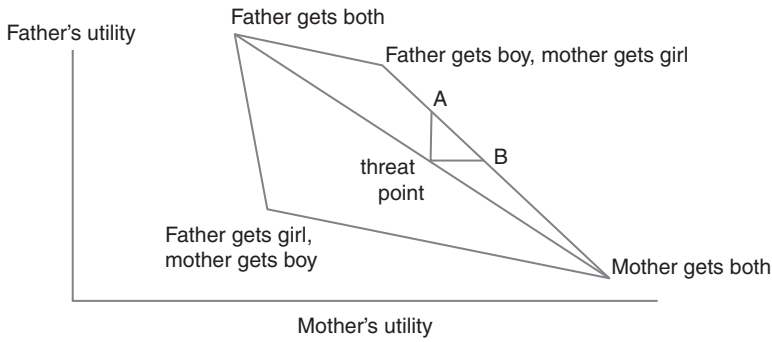



Figure 24.1

line AB, since all combinations of custody and financial settlements on this line have higher utility for both parents than the expected utility of going to court.[26]

Although this analysis gives a rough intuitive understanding of some of the issues involved, it does not capture all of them. In particular, it has no room for inside options, that is, for what happens *during* bargaining and litigation. If one parent has larger financial resources, he (more rarely she) can use them to hire expensive lawyers and expert witnesses. Moreover, if parent A cares more about the harm done to the children by the often painful and protracted custody litigation than does parent B, parent A may be willing to give up custody. Although Solomon would then have accorded custody to parent A, courts cannot take this factor into account.[27]

As the subjective mental states such as impatience and risk aversion that shape the outcome of bargaining cannot be directly observed, bargainers have an interest in misrepresenting them, by verbal or non-verbal behavior. In logrolling, each side will exaggerate the importance of what she is being asked to give up in order to force a large concession by the other. When workers claim to attach great importance to costly safety measures at the workplace, it may be a stratagem to justify a big wage increase as the price of forgoing them. In many cases, attempts to deceive may be too transparent to work. If a divorcing parent claims great concern for getting custody of the children to get a favorable financial settlement, the other parent may be able to document a consistent lack of interest in the children before the marriage began to break down or the recent acceptance of a job that involves a great deal of traveling. A farsighted parent might, however, anticipate this problem and lay the

[26] Various, largely irrelevant mathematical theories yield different predictions as to *which* of these combinations will be chosen.

[27] If it did, and were known to do so, parent B might also be tempted to give up custody.

groundwork for a claim to care about the children before the other parent understands that the marriage is breaking down.

Like parties engaged in arguing, bargainers can have an incentive to misrepresent their interest as based on *principle*. The reasoning behind the misrepresentation is different, though. In arguing, the parties want to prevent the opprobrium of basing their proposals on naked interest. In bargaining, no opprobrium attaches to expression of interest. Firms and workers are supposed to be concerned with profits and wages, not with the common good. Bargainers may nevertheless gain a strategic advantage from framing their demands in terms of principle. They may claim that in backing down from a principle-based claim they are making a greater concession, and hence expect greater concessions from the other side, than if mere interest is at stake. If each side employs this tactic, however, the bargaining may break down.

Negotiations over the allocation of emission rights are vulnerable to this problem, as nations may be attracted to the principles that fit their material interests. One study focuses on four principles:

- *The egalitarian rule* incorporates the principle of equal per capita emissions. It implies that a country whose population amounts to x percent of the global population should receive x percent of global entitlements for greenhouse gas emissions.
- *The sovereignty rule* incorporates the principle of equal percentage reduction of current emissions. It implies that a country whose greenhouse gas emissions amount to x percent of global emissions should receive x percent of global emissions entitlements.
- *The polluter-pays rule* incorporates the principle of equal ratio between abatement costs and emissions. It implies that a country whose greenhouse gas emissions amount to x percent of global emissions should bear x percent of global abatement costs.
- *The ability-to-pay rule* incorporates the principle of equal ratio between abatement costs and GDP. It implies that a country whose GDP amounts to x percent of gross world product should bear x percent of global abatement costs.

The authors of the study first assessed the costs of each of these principles for Russia, the European Union, China, and the United States. They then reported the results of a survey carried out among agents involved in climate policy, asking them to assess, for each of the equity rules, how much these countries or groups of countries could be expected to support it. There was a clear perception that the EU, the US, and Russia supported the equity principles that would impose the least costs on them. For China, the results were ambiguous. Such strategic use of principles *blurs the distinction between arguing and bargaining.*

The distinction is also blurred when it is unclear whether a statement shall be understood as a *threat* or a *warning*. I understand a threat as a statement by A that A will harm B if B does not do X, and a warning as a statement by A that if B does X something bad will happen to B, but not as a result of an action by A.[28] I understand a promise as a statement by A that A will help B if B does X and an assurance as a statement by A that if B does X something good will happen to B, but not as a result of an action by A.

The wage bargaining system that is emerging in China, notably in Guangzhou, illustrates the blurring between threats and warnings. According to Chinese labor law, individual Chinese workers are allowed to strike. At least in theory, they are paid while striking and are not penalized by dismissal for striking. Worker collectives are allowed to form trade unions and elect a leader. However, unlike Western trade unions, their Chinese counterparts are not allowed to threaten to strike.[29] Union leaders get around this obstacle by warning the management that workers are so discontented with their wages or working conditions that they will strike unless their demands are met. They can also refer to highly publicized incidents at other factories, such as Foxconn in Zhenzhen where fourteen workers killed themselves in 2010 in despair over their working conditions, and suggest that similar events might happen at the local plant. These are simple factual statements, to be assessed for their truth or falsity, whereas a threat is assessed by its credibility or lack of it.[30] In reality, of course, what the workers do is to some extent under the control of the union leader, since he is in a position to influence their state of mind.

Another instance of a threat disguised as warning occurred in the Constituent Assembly at Versailles on July 9, 1789, when the Comte de Mirabeau addressed Louis XVI directly after troops were concentrated around the assembly. He first stated that the "French soldiers, close to the center of discussions and sharing the passions as well as the interests of the people, may forget that a commitment made them soldiers, and remember that nature made them men." Technically, the statement was a warning, not a threat. It would have been a threat had he made the assertion – which would have made him liable to a charge of treason – that he would *remind* the soldiers that nature had made them men. Since the speech would be instantly diffused throughout

[28] Some writers use the warning–threat opposition for the distinction between what in my usage are credible and non-credible threats.

[29] In Western countries, too, it is sometimes illegal to threaten (or promise) to do what it is legal to do. The law does not prevent a woman from telling her lover's wife that they have had an affair, but if she threatens to do so unless he pays her off, it is blackmail and illegal. Voters are free to cast their vote for any of the candidates in the running, but are not allowed to promise to vote for one of them in exchange for money.

[30] Western trade union leaders may also disguise their threats as warnings, but for strategic rather than for legal reasons.

the army, it was a *self-fulfilling warning* – not a threat, but close enough. Next, he warned the king that the deputies might lose control of themselves. "We are only men: our distrust of ourselves, the fear of appearing weak, might lead us beyond what we want." Technically, this was a warning too, since he was asserting that the future actions of the deputies would not be under their own control. The effect was the same as that of a threat.[31]

A further example can be taken from the debates of the Federal Convention over the representation of the states in the Senate. The bone of contention was whether all states would have equal representation, as the small states demanded, or whether representation would be proportional to population, as the large states demanded. On June 30, 1787, the delegate Bedford from Delaware, a state that had asserted the equal representation aggressively, claimed that "The Large States dare not dissolve the confederation. If they do the small ones *will find some foreign ally* of more honor and good faith, who will take them by the hand and do them justice. He did not mean by this to intimidate or alarm. It was a natural consequence; which ought to be avoided by enlarging the federal powers not annihilating the federal system."

This was incendiary language, with the reference to "natural consequence" underlying the credibility of the threat. On July 5, Gouverneur Morris from Pennsylvania counterattacked:

> Let us suppose that the larger States shall agree; and the smaller refuse: and let us trace the consequences. The opponents of the system in the smaller States will no doubt make a party and noise for some time, but the ties of interest, of kindred & common habits which connect them with the other States will be too strong to be easily broken. In N. Jersey particularly he was sure a great many would follow the sentiments of Pena. & N. York. This Country must be united. If persuasion does not unite it, the sword will. He begged that this consideration might have its due weight. The scenes of horror attending civil commotion can not be described, and the conclusion of them will be worse than the term of their continuance. The stronger party will then make traytors of the weaker; and the Gallows and Halter will finish the work of the sword. How far foreign powers would be ready to take part in the confusion he would not say. Threats that they will be invited have it seems been thrown out.

The statement can be read as both a warning and a threat. Some delegates certainly took it as a threat, as indicated by the following retreat on behalf of Morris by Williamson from North Carolina: he "did not conceive that Mr. Govr. Morris meant that the sword ought to be drawn agst. the smaller states.

[31] Similarly, Gibbon asserts that a statement by Bishop Ambrosius to the ministers of the Emperor Valentian that "*he* had not contributed to excite, but it was in the power of God alone to appease, the rage of the people; he deprecated the scenes of blood and confusion, which were likely to ensue" could be "interpreted as a serious declaration of civil war." Gibbon also refers to a similarly ambiguous statement (perhaps inspired by Ambrosius) by Cardinal de Retz to Anne of Austria.

He only pointed out the probable consequences of anarchy in the US." In other words, Morris had not made a threat, only pointed out the consequences that could be predicted. On the same day, July 5, Bedford also retreated, by making it clear that:

> he did not mean that the small States would court the aid & interposition of foreign powers. He meant that they would not consider the federal compact as dissolved until it should be so by the acts of the large States. In this case the consequence of the breach of faith on their part, and the readiness of the small States to fulfil their engagements, would be that *foreign nations having demands on this Country would find it in their interest to take the small States by the hand*, in order to do themselves justice.[32]

In Chinese wage bargaining and in Mirabeau's address to the king, the resort to the language of warnings was probably due to the fact that threats would have been illegal and even treasonable. In Philadelphia, the cause may have been the social opprobrium attached to the overt use of threats. Even in this small assembly debating behind closed doors, in which many delegates based their claims on naked interest, threats were beyond the pale.

In bargaining situations, each side may consist of several groups with different aims. Thus in negotiated transitions from authoritarian to democratic political systems, the government as well as the opposition may be divided into hardliners and softliners, the former being more unwilling to compromise. Thus in negotiations between the softliners on the two sides, each may refer to the hardliners in their own camp to argue that there are limits to how much they can concede. They do not, that is, threaten to carry out any particular actions, only warn about what their hardline allies might do. The same principle can apply in international relations. In his memoirs, Richard Nixon wrote that his frustration with Syngman Rhee's tendency to act independently of the United States was assuaged when Rhee told him that "any statements I have made about Korea acting independently were made to help America ... The moment the Communists are certain that the United States controls Rhee, you will have lost *one of your most effective bargaining points* ... The Communists think that America wants peace so badly that you will do anything to get it ... But they do not think that this is true as far as I am concerned, and I believe you would be wrong to dispel their doubts in that respect" (my italics).

## Summary

Pulling together the various strands of this chapter, the process of collective decision making can be represented as shown in Figure 24.2. The central point

[32] Note that in the first statement by Bedford that I have italicized, the initiative of an alliance with foreign allies is imputed to the States, while in the second it is imputed to the foreign nations.

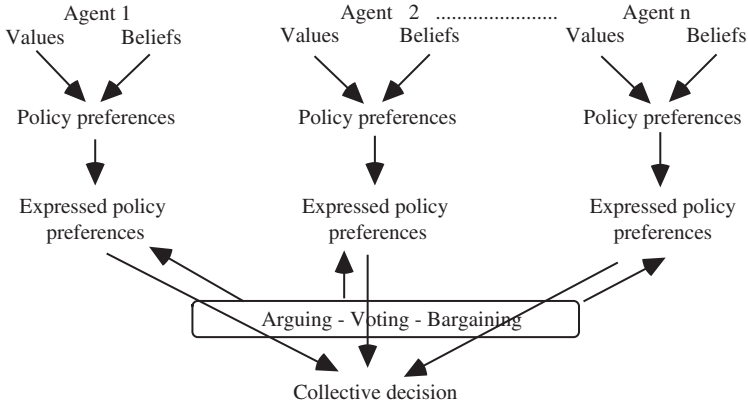



Figure 24.2

is perhaps that each of the mechanisms of collective decision making – arguing, voting, and bargaining – creates an incentive to misrepresent some aspect of one's preferences. In other words, an *aggregation mechanism contributes to shaping the inputs to the mechanism itself*. The expressed policy preferences are a function both of the real policy preferences and of the mechanism that aggregates expressed policy preferences. The welfare impact of misrepresentation is ambiguous. By virtue of the civilizing force of hypocrisy, the effects may be socially beneficial. In other cases, generalized use of this tactic may create a Prisoner's Dilemma type of situation, in which everybody loses.

### Bibliographical note

The experiments showing willingness to sacrifice personal gains for the sake of future generations is described in L. Putterman *et al*., "Cooperating with the future," *Nature* 511 (2014), 220–3. I discuss arguing and voting at greater length in Chapter 2 of *Securities Against Misrule* (Cambridge University Press, 2013), and bargaining in *The Cement of Society* (Cambridge University Press, 1989). For mechanisms that are in some respects intermediate between collective action and collective decision making, see E. Ostrom, *Governing the Commons: The Evolution of Institutions for Collective Action* (Cambridge University Press, 1990). A luminous if occasionally eccentric discussion of arguing and voting is J. Bentham, *Political Tactics* (Oxford University Press, 1999). The passages quoted from Bentham (translated from French) are in the equally interesting *Rights, Representation, and Reform* (Oxford University Press, 2002), pp. 35 and 122. For the practices of the British Wages Council,

see F. Bayliss, "The independent members of the British Wages Councils and Boards," *British Journal of Sociology* 8 (1957) 1–25. The best descriptive studies of arguing (as distinct from normative analyses) are Aristotle's *Rhetoric* and C. Perelman and L. Olbrechts-Tyteca, *The New Rhetoric* (Notre Dame, IN: University of Notre Dame Press, 1969). The Italian practice of secret voting in parliament is explained in D. Giannetti, "Secret voting in the Italian Parliament," J. Elster (ed.), *Publicity and Secrecy in Votes and Debates* (Cambridge University Press, 2015). For misrepresentation induced by deliberation, see Chapter 5 of my *Alchemies of the Mind* (Cambridge University Press, 1999). The paradox named after the Marquis de Condorcet was first stated in his 1785 *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. G. Mackie, *Democracy Defended* (Cambridge University Press, 2003), contains extensive analyses of cycling social preferences, and a claim that almost all alleged examples of cycling in legislatures are based on flawed readings of the evidence. The example of Oslo airport is taken from A. Hylland, "The Condorcet paradox in theory and practice," in J. Elster *et al*. (eds.), *Understanding Choice, Explaining Behavior: Essays in Honour of Ole-Jørgen Skog* (Oslo Academic Press, 2006). The example of the demobilization of American soldiers is taken from S. Stouffer (ed.), *The American Soldier,* vol. II (Princeton University Press, 1949), Chapter 11. The paradox named after the mathematician Poisson was first stated in his 1837 *Recherches sur la probabilité des jugements en matières criminelles et matière civile.* For misrepresentation induced by voting, see M. Balinski and I. Laraki, *Majority Judgment* (Cambridge, MA: MIT Press, 2010). This work also offers an important challenge to the main paradigm of voting theory. For the vote on bicameralism in 1789, see J. Egret, *La révolution des notables* (Paris: Armand Colin, 1950). The discussion of Condorcet's jury theorem draws on D. Karotkin and J. Paroush, "Optimum committee size: quality-versus-quantity dilemma," *Social Choice and Welfare* 20 (2003), 429–41. A full treatment of the history of the secret ballot is H. Buchstein, *Öffentliche und geheime Stimmangabe* (Baden-Baden: Nomos, 2000). The note on the Civil Rights Act of 1964 is taken more or less verbatim from H. Brady and J. Ferejohn, "Congress and civil rights policy: an examination of endogenous preferences," in I. Katznelson and B. Weingast (eds.), *Preferences and Situations* (New York: Russell Sage, 2005). The seminal work on bargaining is T. Schelling, *The Strategy of Conflict* (Cambridge, MA: Harvard University Press, 1960). The argument that bargaining tends to eliminate the gains that are the object of bargaining is in L. Johansen, "The bargaining society and the inefficiency of bargaining," *Kyklos* 32 (1979), 497–522. A classic work on bargaining in practice is H. Raiffa, *The Art and Science of Negotiation* (Cambridge, MA: Harvard University Press, 1982). The claim that the Glorious Revolution enabled English monarchs to make credible promises

is made in D. North and B. Weingast, "Constitutions and commitment," *Journal of Economic History* 43 (1989), 803–32, and critically examined in several chapters in D. Coffman, A. Leonard, and L. Neal, *Questioning Credible Commitment* (Cambridge University Press, 2013). The reference to Louis XV's minister Machault is from M. Marion, *Machault d'Arnouville* (Paris: Hachette, 1891), p. 365. An informal exposition of bargaining theory is A. Muthoo, "A non-technical introduction to bargaining theory," *World Economics* 1 (2000), 145–66. I discuss wage bargaining in Chapter 2 of *The Cement of Society*, and child custody issues in Chapter 3 of *Solomonic Judgments* (Cambridge University Press, 1989). On Robert Walpole and the Sinking Fund, see P. Langford, *Public Life and the Propertied Englishman* (Oxford University Press, 1991), p. 155. On Reagan and Gorbachev in Reykjavik, see K. Adelman, *Reagan at Reykvavik: 48 Hours that Ended the Cold War* (New York: Broadside Books, 2014). For misrepresentation induced by bargaining, see J. Sobel, "Distortion of utilities and the bargaining problem," *Econometrica* 49 (1981), 597–617. For the misrepresentation of interests as principles in climate change negotiations, see A. Lange *et al*., "On the self-interested use of equity in international climate negotiations," *European Economic Review* 54 (2010), 359–75. For some skeptical and commonsensical comments on the importance of the issue of misrepresentation in voting and bargaining, see L. Johansen, "The theory of public goods: misplaced emphasis?" *Journal of Public Economics* 7 (1977), 147–52.

# 25 Institutions and constitutions

### The principal–agent problem

An institution may have *members* or *employees*. Members can also be employees, as in workers' cooperatives. Members interact horizontally, through the processes of arguing, bargaining, and voting that I discussed in the previous chapter. The vertical relations between employees and their superiors have a different character. To simplify, assume that the organization has a single executive ("the principal") and many employees ("the agents"). In a cooperative, the employees collectively are the principal and individually act as agents. A *principal-agent problem* arises when, as is often the case, the principal and the agents have different interests. Workers may have an interest in a moderate work pace, whereas the manager may wish them to make a stronger effort.

Similarly, the head of a state agency has an interest in employees' being honest, in the sense of not taking or demanding bribes from the public. She also has an interest in efficiency, so that the size of the public sector can be kept to a minimum. Employees can have opposite interests in both respects. If they are motivated only by their economic self-interest, they will take bribes when they can get away with it. As a consequence of their interest in power, agents also have an incentive to swell the size of their departments and to multiply the number of subordinates. Again, monitoring may be difficult. The principal may sometimes catch an agent taking a bribe, but in general this is not a method she can rely on. She may try to reduce opportunities for corruption, for instance, by having competitive bids for public contracts, but this precaution does not help if agents tailor the contracts so as to favor particular suppliers. Because agents often have a near-monopoly on information, principals may be unable to tell which requests for more hirings are justified and which are not. During the Cold War, wildly exaggerated estimates about the economic and military power of the Soviet Union produced by the American defense establishment justified a massive build-up of weaponry.

Subordinates are not the only ones that may have incentives out of line with the organization. American university presidents have had to step down when it turned out that they were awarding themselves huge salaries or had their

homes redecorated at the organization's expense. One American vice president (Spiro Agnew) had to resign when he was exposed for corruption. Kleptocracy – the rule of thieves – has become an exceedingly familiar phenomenon across the world. Although they are principals in one sense, such leaders may also be subject to monitoring. Yet overseers (voters, boards of trustees, shareholders, the World Bank, or the International Monetary Fund) have often been conspicuously unsuccessful in regulating the behavior of chief executive officers or heads of state. As in other cases, they may lack either the *information* needed to correct excesses or the *incentive* to do so.

In workers' cooperatives, a conflict may arise between workers as principals and workers as agents. Speaking to the Social Science Congress in 1863 Sir James Kay Shuttleworth said, with respect to the cooperative Lancashire cotton mills:

> [Then] arose the formidable question – What benefits should the shareholders have in this mill beyond the ordinary profits? The first claim that was practically put forward in such societies was, that a preference should be given to the families of the shareholders in selecting the workers in the mill . . . He had witnessed on his own property the failure of one of these concerns. There was a desire to introduce into the concern the principle of cooperation to this extent, that the shareholders should have the advantage of the employment of their families in the mills. The immediate effect of that was this, that instead of producing a stricter discipline and that close attention to the working of machinery, which was so necessary in cotton mills (and he might mention that the discipline of a regiment was inferior in strictness to that of a cotton mill), at their quarterly or half-yearly meetings most vexatious complaints were made by the workers against the overlookers, and an overlooker who had dared to discharge a worker who was a shareholder, was in extreme danger of being dismissed at the next meeting.

Another frequently occurring problem in cooperatives arises from their reluctance to lay off their members in times of low demand. Commenting on the downfall of the Wolverhampton Plate-Locksmiths in 1878, a contemporary wrote that

> if the business had been carried on by a private manufacturer, he would probably have discharged the workmen for whom, from the falling off of the demand for plate-locks, he could not find profitable employment, and applied himself to develop the trade that remained. But this would have involved on the body of workers who formed the society an amount of self-sacrifice for which they were not prepared. Instead, they worked for stock, in the hope that the demand would revive. As it did not revive before their resources were exhausted, they inevitably came to grief. Debts multiplied upon them; the best workers fell away.

There are three main ways in which a principal can respond to opportunistic or otherwise undesirable behavior by the agents: by limiting their *opportunities* to misbehave, by monitoring their *behavior,* and by providing *incentives* that will align their behavior with his.

The first solution is hard to implement. To be effective, an agent needs some independence and freedom of action. Slaves have rarely been used in occupations that demand application and care. It may prove impossible to structure the situation so that the agent is free to pursue the aims of the principal while literally unable to pursue his own. One may try to approximate this goal by having decision makers serve for such a short period that it is hard to bribe them. The jury system and the American electoral college (in its original conception) have been justified on these grounds (among others). Short tenure of elected officials (often combined with non-reeligibility) and frequent rotation of appointed officials, practiced by both the Roman and Chinese empires, are also supposed to reduce the opportunity for corruption. In two compelling metaphors, John Lilburne told the private soldiers in the New Model Army, "Suffer not one sort of men too long to remaine adjutators [delegates], lest they be corrupted by bribes of offices, or places of preferment, for standing water though never so pure at first, in time putrifies," and the Jacobite John Byron described co-opted bodies "as stagnant pools, which needed regular elections to clean them." Short tenure has severe efficiency costs, however. By the time officials have become familiar enough with the job to do it well, they may have to leave it.[1] In many modern governments, the actual and expected turnover of ministers is so high that their time horizon is truncated over and above the normal shortening effect produced by the electoral cycle.

Historically, a common method has been to *monitor* the actions of the agents, not only to prevent shirking, but also to deter theft. Traditionally, that was the task of the foreman, but how can the manager be sure that the foreman is not demanding bribes from the workers or using his authority to promote his own financial or sexual ends? The nineteenth-century workplace was often characterized as "the tyranny of the foreman." In department stores, floorwalkers exercised a similar power. In such cases, who shall guard the guardians? Also, the principal's monitoring of the agent is implicitly based on suspicions that may turn out be self-fulfilling. Monitoring can easily cause resentment, bad morale, and lower productivity. Finally, monitoring is costly. An article from 1989 reported that 7 percent of workers in the US non-agricultural private sector were employed as supervisors or inspectors. (I was unable to retrieve more recent data.)

Relying on incentives may seem more promising, since they do not require the principal to monitor actions, only outcomes. For this reason, incentive

[1] Both the Roman and the Chinese empires adopted two procedures that were not vulnerable to this problem. In China, the "law of avoidance" forbade officials from serving in their native province. To crush a rebellion in one province the governments used troops from another. In Rome, archers on horseback were recruited from Palmyra, but deployed in the Sahara. On June 4, 1989 the Chinese government used a similar stratagem.

systems – rewarding individuals or groups according to their contribution to the goal of the principal – have often been seen as a panacea. The most direct application is observed in firms and institutions that use piece rates. A soccer coach may reward his players according to how many goals they score, or an administration may reward courts according to how many executions they carry out.[2] Norwegian universities at one point gave a bonus to teachers or to their departments for each successful Ph.D. candidate, sometimes with a higher bonus for a female student. Individual teachers may get a salary raise, and departments more positions, as a function of the number of articles published in peer-reviewed journals. The budget of universities may be determined in part by their place on the Shanghai rankings, that of hospitals by the success rate of surgery or the turnover time for patients, and that of police forces by the number of reported crimes or by the percentage of crimes that are solved. Some organizations use tournaments (winner-takes-all competitions) among their members, notably to determine who shall be promoted. Examples could be multiplied indefinitely.

Institutions can also use *negative incentives* to induce compliance with the principal's goals. The boot camp is a paradigmatic example. In the classroom, positive as well as negative incentives are used. In France, students are (or were) moved around the classroom every week according to their performance, good students being moved to the front and less good to the back, thus combining the carrot and the stick. In a field experiment, students who earned their grades ("earners") received positive incentives and those who maintained their grades ("maintainers") received negative incentives. The earners started the semester with 0 points and added points with each graded assignment, whereas the maintainers were given the maximum number of points available for the course at the beginning of the semester and then subtracted points from this overall total as they lost points on a graded assignment. Consistent with the theory of loss aversion (Chapter 14), the maintainers performed (slightly) better.

An incentive system can succeed in aligning the interests of the principal and the agent or agents, as in the following stylized example. Suppose that the principal has an asset whose value may be high or low depending on whether the effort of the agent is high or low. The principal is a mayor who runs for reelection on the promise of a successful public safety campaign. The likelihood of reelection is high if the agent, the chief of police, supplies a high effort rather than a low effort. Since effort is costly to the agent, it has to be elicited

[2] According to Gibbon, under the reign of Valens and Valentian judges "easily discovered that the degree of their industry and discernment was estimated, by the Imperial court, according to the number of executions that were furnished from their respective tribunals." In contemporary societies, *quotas* seem to be the more common pathology of legal incentive systems.

by incentives. Assume that the high and low values of the asset are, respectively, \$30,000 and \$10,000. The costs to the agent of a high and low effort are, respectively, \$8,000 and \$4,000 (he has to make *some* effort to avoid being fired[3]). The probability of the high-value outcome being realized with high effort is 80 percent; with a low effort it is 30 percent. Suppose the principal proposes a bonus B in case she is reelected. If the agent makes a low effort and the mayor is reelected, his benefit is 0.3B – 4,000. (Since the principal can observe only the outcome, not the action, the agent might get the bonus even if he is slacking.) If he makes a high effort, the benefit is 0.8B – 8,000. A simple calculation shows that the agent's benefit for hard effort exceeds that of low effort when the bonus is greater than \$8,000. If the principal pays the bonus, her net benefit is \$18,000 (0.8 × \$30,000 + 02 × \$10,000 – \$8,000). If she does not, it is \$16, 000 (0.3 × \$ 30,000 + 0.7 × \$10,000). Hence she has an incentive to pay the bonus, which will give the agent an incentive to work hard. Both parties gain.

In this case, there was only one principal and one agent. In many important cases, a principal has to choose one of several agents. If each of the latter has private information about the facts on which the principal will base his decision, her task is to create incentives for the agents to reveal the information, not verbally, but through the choices they make. An early, if imperfect example is provided by the judgment of Solomon:

> Then came there two women, *that were* harlots, unto the king, and stood before him. And the one woman said, O my lord, I and this woman dwell in one house; and I was delivered of a child with her in the house. And it came to pass the third day after that I was delivered, that this woman was delivered also: and we *were* together; *there was* no stranger with us in the house, save we two in the house. And this woman's child died in the night; because she overlaid it. And she arose at midnight, and took my son from beside me, while thine handmaid slept, and laid it in her bosom, and laid her dead child in my bosom. And when I rose in the morning to give my child suck, behold, it was dead: but when I had considered it in the morning, behold, it was not my son, which I did bear. And the other woman said, Nay; but the living *is* my son, and the dead *is* thy son. And this said, No; but the dead *is* thy son, and the living *is* my son. Thus they spake before the king. Then said the king, The one saith, This *is* my son that liveth, and thy son *is* the dead: and the other saith, Nay; but thy son *is* the dead, and my son *is* the living. And the king said, Bring me a sword. And they brought a sword before the king. And the king said, Divide the living child in two, and give half to the one, and half to the other. Then spake the woman whose the living child *was* unto the king, for her bowels yearned upon her son, and she said, O my lord, give her the living child, and in no wise slay it. But the other said, Let it be neither mine nor thine, *but* divide *it*. Then the king answered and said, Give her the living child, and in no wise slay it: she *is* the mother thereof. (King James Bible, *1 Kings* 3)

[3] That minimal effort has to be observable. This seems realistic: it is easy to observe whether an agent performs routine duties, but not whether he shows initiative.

The example is imperfect, in a technical sense, because the false claimant has an incentive to mimic the true one, leaving Solomon with no useful information.[4] The modern theory of incentives tries to overcome this problem, by creating mechanisms that will *separate* true from false claimants, or "good types" from "bad types." As shown in one of the first analyses of this issue, employers may use years of education as a signal of productivity – *not* because education makes people more productive, but because the cost of undertaking an education is lower for more productive individuals. Anecdotally, the British civil service was run on this principle: if a person can acquire the capability in three years to compose Greek and Latin verse, he (very rarely she) was also qualified to run part of the empire. Applicants *sort themselves out* into "good types" and "bad types" by their willingness to undertake education. Similarly, low-risk drivers can signal their type by choosing insurance policies with higher deductibles but lower premiums, since their risk of having to pay the deductible is less than that of high-risk drivers.

Sometimes, as I shall now discuss, incentive systems fail. There are three main sources of failure: the incentives may induce rational behavior contrary to the intended one; they may change the preferences of the agents, undermining the standard (if usually tacit) assumption that preferences will be unaffected by the scheme; and they may induce irrational behavior that undermines the equally standard assumption that people respond rationally to incentives.

Individual incentives may come at the expense of group performance. If individual soccer players are rewarded by how many goals they score, they may prefer to make a shot from a long distance or at a sharp angle rather than passing the ball to someone who is in a better position to score. Since the group is small, the members may monitor one another. They might punish a player who does not pass the ball to players who are in a position to score by not passing it to him when he is in that position. The use of team bonuses might also, as in the case of waiters who pool their tips (Chapter 23), create an incentive to monitor players who do not pull their full weight. In other cases, neither of these corrective mechanisms may be available. In an incentive-based competitive environment, agents may not be willing to share their information with each other even though the principal would benefit if they did.

Incentive systems can *backfire* if they create a fear of a moving target. If a system succeeds in making employees work harder, they may fear that the higher-level effort will be the new benchmark. As noted in Chapter 23, this may lead to the emergence of norms against rate busting. In France, research fellows in the social sciences and the humanities at the National Council of Scientific Research (CNRS) have reacted in similar ways against what they

[4] Similarly, the norm that was prevalent in eighteenth-century America, that seeking political office was a sign of being unworthy of it, left voters with no useful information.

claim to be an excessively individualistic and result-oriented approach. Although they have life tenure and no teaching obligations, the outcome of their grant applications may depend on the evaluation of their research; hence high standards are a threat. As is the case in the same disciplines in the Italian academic system, these sections of the CNRS are permeated by a *norm against excellence*.

A more general and more important problem is that many – some would say *all* – incentive schemes create an opportunity for *gaming the system*, in ways that range from cheating to exploiting loopholes.[5] Let me cite some examples.

- A famous caricature in the Soviet magazine *Krokodil* showed a nail factory that fulfilled its quota (specified in weight) by producing one gigantic nail.
- If you cannot win the promotion tournament by your own effort, you can always trip up the competition.
- A transplantation center that is judged by the success of its grafts can increase its budget allocation by refusing to accept patients with a bad antigen match.
- In a British hospital, patients who should have waited no longer than four hours before they were seen were left sitting in ambulances because the clock did not start ticking until they entered the building.
- Some British hospital managers inquire when patients intend to go on holiday and then offer an appointment during this period. Few patients cancel their holiday for medical reasons, preferring to postpone their appointment. Since the patients initiate these delays, their wait is not recorded.
- "Coding creep": if a major risk factor is recorded in a higher proportion of patients before surgery, the unit's predicted mortality will increase, as will the likelihood that the unit's actual mortality falls within or below the expected range.
- When British schools were evaluated by the number of students who obtain good results in their exit exams, they responded by excluding more students, which led to more local petty crime.
- The system of rewarding scholars by their peer-reviewed publications has created the concept of the "smallest publishable unit," that is, the smallest part of an article that can be taken out and published separately.
- When school boards use student test scores to punish schools, the effect may be, as one study found in Chicago, that teachers falsify the scores they report rather than making an effort to improve the performance of students.

[5] Providing agents with *information* can also have perverse effects. When American officers in Vietnam provided information about the location of Viet Cong soldiers to the South Vietnamese army, they sometimes used it to go where there were no guerrillas.

- When teachers or departments are rewarded for how many Ph.D. degrees they confer, academic standards, like water seeking the lowest level, tend to fall. The No Child Left Behind Act had similar perverse effects.
- The New York police department reclassified some felonies as misdemeanors because the latter did not count in the crime statistics.

These examples, which again could be multiplied indefinitely, confirm the saying that performance targets are good servants, but bad masters.[6]

A much-debated question is whether the use of positive incentives by the principal can *change preferences* by undermining the intrinsic motivation of the agents. Although this effect plausibly operates in some contexts – one should not pay one's children to make their beds or do their homework – it is not clear whether it matters in an institutional setting. *Negative* incentives may certainly change the preferences of the targeted individuals. This effect is a small-scale version of the "psychology of tyranny" that I discussed in Chapter 2. People can resent negative incentives for the same reasons for which they resent monitoring. In the short run, compliance is increased; in the long run, initiative suffers.

Finally, incentive schemes may fail if the agent does not respond rationally, for one of the many reasons uncovered by behavioral economics. The self-sorting of individuals into high-risk and low-risk drivers is undermined if, as seems to be the case, a large majority of people believe they drive better than the average person. Although pay-for-performance incentives are widely used in the health sector, they have had little impact. The reason may be that programs are poorly designed and do not reflect what is known about the psychology of how people – including doctors – respond to incentives. One study examined a program designed to increase the number of women who receive a mammogram, and identified seven psychological mechanisms that may have contributed to its poor success rate. One failure was that the incentive payment was an incremental increase of the usual reimbursement, e.g. increasing the reimbursement per visit from $100 to $106. The designers of the scheme neglected the minimal psychological effect of this change, due to an irrational tendency to underestimate the difference between $100 and $106.[7] The authors also make the ironic observation that attempts to incorporate elements from behavioral economics could actually backfire. As I noted in the discussion of the use of negative and positive incentives in the classroom,

[6] Organizational failures can also arise if officials fall victim to what I called "the younger sibling syndrome" (Chapter 13) and neglect the fact that people respond to incentives. This is a cruder mistake than neglecting the fact that they can respond to incentives in more than one way.

[7] A more vivid example: people who would normally weigh carefully the quality difference between stereo players costing $100 and $150 usually do not care about the difference between a $20,000 car with the $100 player and the same car with the $150 player.

the former may be more effective because of loss aversion. Since incentive payments to doctors can also be structured either as a withholding (a perceived loss in income) or as a bonus (a perceived gain), performance should improve by using the first scheme. As in other cases of negative incentives there is a risk, however, that this advantage might be outweighed by the risk of angering physicians.

## Constitutions

*Constitutions* can be studied from two perspectives, to understand their upstream causes or their downstream effects. I begin with some comments on the former.

The making of constitutions can be a collective action problem (Chapter 23), notably if it requires different social groups to agree on a common tax policy. In the United States before 1787 and in France before 1789, each state or each estate wanted to benefit from the public goods – infrastructure, law and order, national defense – funded by taxes, while contributing as little as possible. In America, many states refused calls from the Continental Congress to pay contributions to the common cause; in France, the nobility and many members of the Third Estate sought and obtained tax exemptions. After considerable turmoil, the constitutions that came into effect in, respectively, 1789 and 1791 imposed centralized tax structures.

The Federal Convention overcame another collective action problem, by establishing in the constitution that "Senators and Representatives shall receive a Compensation for their Services, to be ascertained by Law, and paid out of the Treasury of the United States." Under the Articles of Confederation, travel and accommodation expenses of representatives to the Continental Congress were paid by the individual states, which often sent either no representatives or two (the minimum size of a delegation) to save money. Since a two-member delegation could not cast a vote if the two disagreed, the result was often paralysis, as Thomas Jefferson wrote to George Washington on March 15, 1784: "I suppose the crippled state of Congress is not new to you. We have only 9 states present, 8 of whom are represented by two members each, and of course, on all great questions not only an unanimity of States but of members is necessary. An unanimity which never can be obtained on a matter of any importance. The consequence is that we are wasting our time & labour in vain efforts to do business. – Nothing less than the presence of 13 States, represented by an odd number of delegates will enable us to get forward a single capital point."

Constitution making also illustrates the variety of motivations that can animate social agents, and notably the interplay among interest, passion, and reason (Chapter 4).

Passions arise from the fact that constitutions tend to be written in times of crisis and turbulence. The conditions include war (Norway and France 1814, Germany, Czechoslovakia, and Poland after WWI, Germany, Italy, Japan, and France after WWII), revolution (France and Germany 1848), the fall of a dictatorship (Greece 1974, Portugal 1974, Spain 1976, South Africa 1996, many Latin American countries in the 1980s), the fear of a coup (France 1958), regime implosion (Eastern Europe after 1989), and financial crises (US 1787, France 1789, Hungary 2010, Iceland 2011). As I noted in Chapter 8, the emotions that are triggered include *fear* and *enthusiasm*. In Iceland, the predominant emotion was *anger* at the banks, which were perceived to be responsible for the financial crisis.

Reason – the impartial and rational concern for the long-term public interest – is both facilitated and hindered by passion. In Chapter 4 I cited La Bruyère, "Nothing is easier for passion than to overcome reason; its greatest triumph is to conquer interest." One could add, as a corollary, that nothing is easier for interest than to overcome reason, *except when reason allies itself with passion*. As Kant noted, enthusiasm can produce this alliance. As he also observed, however, passion can be an obstacle to rational belief formation (see Chapter 8). By the mechanisms of wishful thinking and urgency, enthusiasm can lead agents to ignore obstacles to their goals that cool and deliberate reflection would have made obvious. In Iceland in 2011, the urge to *do something* in response to the financial crisis preempted the question whether making a wholly new constitution was the best thing to do. Framers may even ignore their own reason-based decisions, as when both the French framers in 1789 and the Norwegian framers in 1814 ignored their own rule that proposals should be subjected to several successive votes before being adopted. As Gibbon says about the measures taken to enforce the residence of elected bishops in the reign of Constantine, "the same passions which made those regulations necessary, rendered them ineffectual."

Even though the alliance of passion and reason may override some interests, they are never completely eliminated. In modern constitution making, *party interest* is often decisive in shaping electoral laws. In earlier periods, *class interest* often shaped the limitations on suffrage and eligibility. At the Federal Convention, the *economic interests* of the Southern and Northern states shaped at least a dozen clauses in the final document. In a few cases, *personal interests* have played a role. The creation of a Senate in the Czech constitution of 1992 was not made on grounds of principle, but to create a place for the Czech deputies to the Senate of the dissolved Czechoslovak Confederation.

I now consider the structure and effects of constitutional texts. A constitution establishes the separation of powers. It prevents any single political actor from concentrating all power in its hands. The single actor may be an individual, a small group, or the people as a whole. The classical

terms for their unconstrained power are tyranny (absolute monarchy), oligarchy, and mob rule.[8] When constrained by the separation of powers, the regimes turn into constitutional monarchy, aristocracy, and democracy. Here, I consider only modern democratic constitutions. Typically, these have four parts. First, they determine and regulate the machinery of government. Second, they specify the rights and sometimes the duties of the citizens. Third, they lay down rules for amending the constitution. Finally, they stipulate procedures for suspending the constitution, or specified parts of it, in times of an emergency.

The machinery of government has many nuts and bolts, as well as cogs and wheels. The core institutions are the legislative, executive, and judicial organs. As Bentham argued, one should also include the electorate as an organ. Other important institutions include the national auditing office and the Central Bank. Many provisions relative to these organs are simply lists of their functions: the power to coin money, to raise taxes, to increase the money supply, to sign treaties, to enact legislation, to decide in civil and criminal cases, to vote in elections, and so on.

Other provisions involve relations to other organs. One may, perhaps, distinguish two bundles of such relations. One bundle aims at preventing one organ from trespassing on the domain of another. In Sweden, the government cannot instruct the Central Bank in matters of monetary policy. In some countries it is also prohibited from instructing the public prosecutor. A ban on bills of attainder prevents the legislative power from encroaching on the judiciary. Conversely, in France the fear of a "government of judges" blocked the judicial review of legislation for two centuries. The other bundle consists of mutual checks among these organs. An organ can check another to prevent it from making formally unconstitutional decisions, or members of one organ can check another to prevent it from making what they think (or claim) to be substantively bad decisions. Judicial review and bicameralism illustrate these two forms.[9]

[8] Objecting to the tendency to blame the sovereign for all disorders, Hume wrote that "As if the turbulence of the great, and madness of the people, were not, equally with the tyranny of princes, evils incident to human society, and no less carefully to be guarded against in every well regulated constitution."

[9] The second bundle cannot be neatly distinguished from the first. For one thing, if organ A checks organ B on the grounds that organ B has acted unconstitutionally, the act may be one that encroached on the constitutional powers of organ A or of some other organ C. For another, the dividing line between checks and encroachments is relative. The right of the American Senate to veto the appointment of high federal officials is usually presented as a desirable check on the executive, whereas the right (sometimes claimed in the past) also to veto their removal has been seen as an inappropriate encroachment. In other systems, even the first right would be seen as an encroachment. Judicial review can be assessed either negatively as an encroachment on the domain of the legislature, or positively as a check on its activities. This ambiguity is pervasive. For those who follow Bentham in thinking that the lower house of parliament should be omnipotent, *any* check on its power is an encroachment. His view is at one extreme of a continuum. The American constitution with its triple check on the lower house – by the upper house, by the president, and by the Supreme Court – may be at the other extreme.

All modern constitutions (except Australia's) include a bill of rights, or at least an enumeration of rights. Civil and political rights are usually defined vertically: they forbid the government from taking certain actions against the citizens. Only in South Africa is there an express commitment to the horizontal application of the constitution. In addition to these "first-generation rights," modern constitutions increasingly include "second-generation rights" (economic, social, and cultural rights) and the more diffuse "third-generation rights" (e.g. the right to development). Whereas constitutional articles related to the machinery of government are (as a general rule) formulated so sharply that it is unambiguous what can, what cannot, and what must be done, articles affirming rights acquire (again as a general rule) implications for action only when filtered through statutory law or constitutional jurisprudence. I shall return to this question.

Virtually all constitutions provide procedures for their own amendment. Tocqueville observed that the French Charter of 1830 did not contain an amendment clause, from which fact he drew the conclusion that judicial review – "le gouvernement des juges" – would be too dangerous: "If courts in France could disobey laws on the grounds that they found them unconstitutional, constituent power would really be in their hands, since they alone would have the right to interpret a constitution *whose terms no one else could change*." There may have been other wholly unamendable constitutions, but I doubt they are important. That being said, some constitutions contain individual articles that are unamendable.

If common sense suggests the need to be able to amend the constitution, it also suggests the need to make amendment relatively difficult. If the constitution were as easy to amend as ordinary laws, one might adopt an unconstitutional law by first changing the constitution, to make it constitutional, and then adopt the law. Generally speaking, however, legal systems do not enable agents to do in two steps what they are forbidden to do in one.[10] With insignificant exceptions, it is in fact always more difficult to amend a

[10] Violations of this principle do occur. "In the twenty-third of Henry VI. a law ... was enacted, prohibiting any man from serving in a county as sheriff above a year, and a clause was inserted, by which the king was disabled from granting a dispensation. Plain reason might have taught, that this law, at least, should be exempted from the king's prerogative: But ... in the reign of Henry VII. the case was brought to a trial before all the judges in the exchequer-chamber; and it was decreed, that, notwithstanding the strict clause abovementioned, the king might dispense with the statute: He could first, it was alleged, dispense with the prohibitory clause, and then with the statute itself. This opinion of the judges, though seemingly absurd, had ever since passed for undoubted law" (Hume). Two other instances were the decision by Tiberius (as reported by Tacitus) to kill the daughter of his enemy Sejanus and the execution of young women in the Iran of the Ayatollahs (as reported in the *Boston Globe* of August 19, 2009). In both cases, it was illegal to execute virgins; in both, the problem was circumvented by first raping them.

constitution than to change ordinary laws. *Delays* and requirements of a *supermajority* are the most important devices. A delay means that the minimum time between the proposal of an amendment and its adoption is longer than in the case of ordinary legislation. To achieve this end, the constitution may require that the amendment be debated no earlier than one month from its introduction (Bulgaria), that it be subject to two readings and two votes (Brazil), that it be proposed in one parliament and adopted in the next one after a general election (Belgium, Norway), or that it be passed by two successive parliaments (Denmark, Estonia, Finland, Iceland, Sweden). Supermajorities range from three-fifths to three-quarters, with a two-thirds requirement being perhaps the most common. In Finland, Estonia, and Bulgaria, there is a trade-off between the length of the delay and the size of the supermajority. In the first two countries, there exist *fast-track procedures* by which the normal time-consuming process can be bypassed if a large majority (five-sixths in Finland and four-fifths in Estonia) declares the need to revise the constitution urgently and another majority (two-thirds in both countries) then votes to amend it. In Bulgaria, there exists a *slow-motion procedure* by which the normal supermajority of three-quarters can be reduced to two-thirds but with a longer delay. The Norwegian constitution may also be seen as expressing a trade-off, since ordinary revisions require both a delay and a supermajority of two-thirds, whereas the delegation of certain powers to an international organization requires a supermajority of three-quarters but no delay.

Just as constitutions regulate their own amendments, they sometimes regulate their own partial suspension during emergencies. The suspension may concern one or more of the other three parts of the document, regulating the machinery of government, individual rights, and the amendment process. In addition to determining the *scope* of the suspension, the constitution may also identify the *grounds* for suspension, the *organ* that decides on the suspension, and the procedures for *ending* the suspension. An important difference between amendments and suspensions of the constitution is that the latter may lack a basis in the constitution itself. The difference is not absolute: in 1962 de Gaulle tacitly amended the French constitution by the unconstitutional means of a referendum. Yet in cases of *force majeure,* bypassing the constitution is more common. The suspension of the French constitution of 1793 immediately upon its enactment, in favor of a revolutionary government, had no basis in the document. Saint-Just argued that "In the circumstances in which the Republic finds itself, one cannot establish the constitution; *one would destroy it through itself.*" Another famous example is Lincoln's suspension of habeas corpus in 1862, a decision that the constitution vests in Congress. His response to criticism – "Are all laws but one to go unexecuted, and the Government itself go to pieces lest that one be violated?" – was an

echo of Saint-Just and in turn was echoed in Justice Robert Jackson's dictum, "The bill of rights is not a suicide pact."

This overview provides the background for the question that is most relevant for the purposes of the present book: *how does a constitution acquire causal efficacy?* Would it not simply be a "rope of sand" (Cromwell) or a "parchment barrier" (Madison)? Why would the government behave like the proverbial chicken that stays inside the chalk circle it could easily transgress? According to the authors or inspirators of the 1799 French constitution, "Il faut qu'une constitution soit courte et obscure. Elle doit être faite de manière à ne pas gêner l'action du gouvernement." ("A constitution should be short and obscure. It should be written so as not to interfere with the action of the government.") In a modern perspective, of course, a main function of a constitution *is* to constrain the government. The question is how it can do so.

Actually, the question is too narrowly framed. A constitution does not merely serve the end of disabling the government in some respects, but also that of *enabling* it. A government in a country without a rigid (hard-to-amend) constitution will not be able to make credible promises. This was a major problem for the British parliament in the eighteenth century, when it was unable to make credible promises to the American colonies that it would not tax them in the future. In 1776, James Cannon wrote that "I call upon you to prove that Great-Britain can offer any plan of constitutional dependence which will not leave the future enjoyment of our liberties to hope, hazard, and uncertainty . . . By the constitution [*sic*] of Great-Britain the present Parliament can make no law which shall bind any future one . . . Is it wisdom, then, or is there safety, in entering upon terms of accommodation with a power which cannot stipulate for the performance of its engagement?" A rigid constitution can also enable the citizens to engage in long-term economic planning by removing the chilling fear that the government might confiscate their gains. Yet these observations, while correct, presuppose that the constitution is not only hard to amend, but also *enforceable* and *credible*. The question is how it can acquire these properties.

Before the near-universal adoption of judicial review to prevent unconstitutional actions by the legislature or the executive, various other methods were proposed, adopted, or observed. Cromwell proposed to give the executive the right to veto encroachments on its powers. Some deputies to the French constituent assembly of 1789–91, notably Mounier and Mirabeau, proposed to give the executive the right to veto all laws. In the nineteenth century, France (de facto) and Sweden (de jure) gave the president of the legislatures the right to prevent proposals they judged unconstitutional from coming to a vote. Each of these solutions amounts to making one part judge in its own cause. The British legal theorist Dicey, commenting on the lack of control of

the constitutionality of laws in France around 1900, wrote that the "restrictions placed on the action of the legislature under the French constitution [of the Third Republic] are not in reality laws, since they are not rules which in the last resort will be enforced by the courts. Their true character is that of maxims of political morality, which derive whatever strength they possess from being formally inscribed in the constitution and from the resulting support of public opinion."

That strength, however, may not be very great. There is evidence that citizens worry little about unconstitutional proposals if they approve of their substance. Commenting on the use of royal prerogatives that were widely seen as contrary to the (unwritten) British constitution, David Hume observed that:

> In 1662, Charles, pleading both the rights of his supremacy and his suspending power, had granted a general indulgence or toleration; and, in 1672, he renewed the same edict: though the remonstrances of his parliament obliged him, on both occasions, to retract; and, in the last instance, the triumph of law over prerogative was deemed very great and memorable. In general, we may remark that, where the exercise of the suspending power was agreeable and useful, the power itself was little questioned: where the exercise was thought liable to exceptions, men not only opposed it, but proceeded to deny altogether the legality of the prerogative on which it was founded.

The statement is confirmed by de Gaulle's unconstitutional procedure in 1962, when he changed the constitution by referendum to make the president elected directly by the people. A large majority of the voters voted for the proposal rather than voting against it to punish him for the choice of procedure. Those who voted No mostly did so on grounds of substance, not of procedure. Another case of unconstitutionality in French politics occurred on May 29, 1849, when President Louis Bonaparte gave orders for French troops to march on Rome for the purpose of defeating the Roman Republic and restoring the papacy. The action was clearly a violation of the constitution of 1848, which says that "The French Republic . . . never uses its troops against the freedom of any people." In the National Assembly, the radical deputy Ledru-Rollin stated on June 11 that "The constitution has been violated; we shall defend it by all means, even by arms," to which the President of the Assembly Dupin replied, "The constitution cannot be violated in a more scandalous manner than when in a legislative assembly one talks about defending it by arms." When Ledru-Rollin called for an armed demonstration on June 13, only a few thousand unarmed persons showed up and were quickly dispersed by force.

The modern solution to this problem is to assess the constitutionality of laws by judicial review entrusted to constitutional courts or supreme courts. The existence of a court with the uncontested power to exercise judicial review does not, however, resolve the question of the causal efficacy of the written constitution. Like the constitution itself, *a decision by the court is just a piece*

*of paper.* Courts exercising judicial review usually do not have a police force at their disposal to enforce their decisions. (Nor, for that matter, do they have a separate budget to fund decisions with important economic consequences.) Although the US Supreme Court can ask the President to send federal marshals to enforce its decisions, he can refuse, as did President Jackson in *Worcester* v. *Georgia* (1833). In an apocryphal story, he is reported to have said, "Well, John Marshall [the Chief Justice] has made his decision, now let him enforce it." He did say, though, that the decision was "still born." The decision in *Brown* v. *Board of Education* (1954) would not have been enforced had President Eisenhower not sent federal troops to Little Rock. Undercompliance with decisions by the constitutional court has also been documented in South Africa and in Russia.

In the American cases, opposition to the Court was rooted in the conflict between the states, particularly the Southern states, and the federal government. It might seem hard to imagine similar attempts to sabotage the decisions in, say, contemporary France or Norway. In the case of a priori review of legislation, that is, review of legislation before it is promulgated, they are virtually unthinkable. In the case of a posteriori review, that is, review that arises out of a specific legal case, decisions that entail large expenditures by the authorities might be met with the answer, made in good or bad faith, "We don't have the money." Although the enforcement of economic, social, and cultural rights might be particularly expensive, the enforcement of civil and political rights can also require substantial outlays. For this reason, the government might, for instance, refuse to comply with court-ordered measures to reduce prison overcrowding.

The last example can be used to introduce another reason why the constitution may lack causal efficacy. Reacting to the arbitrary practices of the courts of the *ancien régime*, the *parlements*, Montesquieu said that judges "are no more than the mouth that pronounces the word of law (*la bouche de la loi*), mere passive beings, incapable of moderating either its force or rigor." One interpretation of his statement is that the law, including the constitution, always has a unique meaning that is to be determined by the judge. Some constitutional provisions no doubt come close to having this character, notably those that specify the machinery of government. These clauses provide hard constraints on governmental action. Provisions stating individual rights, by contrast, are notoriously ambiguous. The Eighth Amendment to the American constitution prohibits the use of "cruel and unusual punishment." One might ask whether "double-celling" in prisons – two prisoners sharing one cell – is banned under this clause. In *Rhodes* v. *Chapman* (1981), a majority on the US Supreme Court found that it was not, Justice Marshall dissenting.

It seems crystal clear to me that the Eighth Amendment by itself provides no unambiguous resolution of the "double-celling" issue. The majority and the

minority formed their opinions on the basis of a vast body of prior jurisprudence, most of which has an equally tenuous relation to the text of the constitution, and on their personal ideas of what constitutes unusual and cruel punishment. Although both the majority and the minority cited an earlier decision affirming that the state cannot impose punishment that violates "the evolving standards of decency that mark the progress of a maturing society," they drew opposite conclusions from that shared premise. In *Roe* v. *Wade* (1971), the Supreme Court very dubiously found a right to privacy in the constitution and a right to abortion in the right to privacy. The idea of "substantive due process" has equally poor foundations in the text of the constitution, as do the Supreme Court decisions about what counts as "speech" in the First Amendment and about exactly which right to bear weapons follows from the Second Amendment (see Chapter 9).

Many similarly creative decisions could be cited from other countries. Here, I consider two pairs of decisions by the Colombian Constitutional Court. First, consider the contrast between decision C-221/1994 allowing personal drug consumption and decision C-309/1997 allowing the mandatory use of safety belts. The question in both cases was whether the law could restrict personal autonomy for the purpose of protecting individuals against themselves. In the first decision the Court affirmed that it could not, in the second that it could. The argument in the first decision was that if the state wanted to reduce drug consumption, it should use the less restrictive means of education. This idea is so removed from reality that one must conclude that the Court was in the grip of an ideology. Could not drivers also be "educated" to use safety belts?

Second, consider the contrast between decision C-1040/2005, which authorized a second term for President Uribe, and decision C-141/2010, which denied him a third term. In the first decision, the Colombian Court distinguished between "amendment" of the constitution and "substitution" of one constitution for another. While it denied Congress the right to substitution, it accepted its right to amendment and judged that the extension of the one-term limit was only an amendment. In its 2010 decision, it struck down a law calling for a referendum on an amendment to the constitution for a two-term extension. On the Court's English website, the two decisions are contrasted in these terms: "The [2005] decision ... reviews, in great detail, all the procedural aspects of the amendment's transit through Congress. But its most important feature is the way in which it applied, to a specific case, the 'substitution theory'. Five years later, confronted with another constitutional amendment that allowed for a third consecutive term, the Court decided against it. Two terms is not a substitution, but three terms is." The distinction seems arbitrary. The motivation behind the decision was probably the desire of the judges to keep Uribe out of power.

As these examples illustrate, because of the abstract and often vague language of many rights provisions in the constitution, they can indeed be "adapted to our concerns," as Montaigne said (Chapter 9). *The conclusion generates the premises*. To the (unknowable) extent that this is the case, the constitution has no independent causal efficacy.[11] Even when it is not the case, and judges do their best to decide cases on their merits rather than twist the premises to reach a predetermined conclusion, they are much more constrained by prior constitutional jurisprudence than by the constitution itself. Moreover, as I noted, that prior jurisprudence itself had often very tenuous links with the original text. We can easily imagine a counterfactual world in which many key decisions were made differently, because of the death and replacement of swing judges, with the cumulative effect of creating a constitutional jurisprudence wholly different from the one under which we live today.[12] The constitution itself, though, would be the same.

So far I have focused on the causal efficacy of the rights provisions in the constitution. Some provisions regulating the machinery of government definitely have causal efficacy, certainly by affecting opportunities and perhaps by affecting desires (see Chapter 10 for this distinction).

Regarding *opportunities*, the Twenty-Seventh Amendment, proposed in 1789 and adopted in 1992, states that "No law, varying the compensation for the services of the Senators and Representatives, shall take effect, until an election of Representatives shall have intervened." Before the adoption of that amendment, self-dealing by members of congress was kept in check by their desire to be reelected, together with the belief that voting themselves a salary increase might thwart that desire.[13] If the constitution does not allow the executive to dissolve parliament and call for new elections, it removes an arrow from its quiver.

Regarding *desires*, it is a cliché that a bicameral system will favor wise decisions, by virtue either of the superior virtue and ability of the senators or by the sheer fact of slowing down the process to let emotions cool off. I do not know of any empirical analyses of this alleged effect, but some momentous American examples suggest doubt. In 1798, the Sedition Acts passed the Senate by a wide margin, but obtained a bare majority of 44 to 41 in the

[11] Even the interpretive norm of respecting "the plain meaning of the text" does not always give a clear-cut answer. In twenty-one Supreme Court cases decided with an opinion during the spring of the 1993 term, conflict between majorities and dissents derived at least in part from *disagreements over the plain meaning* of the statute at issue.

[12] Until 1941, swing votes (decided five to four) were rare. After that date, about 16 percent of the cases have been decided with the majority of one vote.

[13] In 1816, there was such a high degree of citizen indignation when legislators voted themselves a pay increase that almost two-thirds of them failed to be reelected, even though they had hastily repealed the compensation law in the meantime.

House of Representatives, and then only after trial by jury was substituted for trial by judge. Many legal scholars think that the Acts would have been found unconstitutional if judicial review had been established at the time. In 1964, the Resolution of the Gulf of Tonkin passed the House by 416 votes to 0, and the Senate by 88 votes to 2. The "Authorization for Use of Military Force Against Iraq Resolution of 2002" passed the House by 297 votes to 133 and the Senate by 77 votes to 23. In none of these cases did the Senate show much resistance to the whipped-up atmosphere of hysteria. These three decisions are usually viewed as embodying the opposite of wisdom. Since the American Supreme Court is also supposed to act as a brake on impetuous decisions, it is worth mentioning that it approved Roosevelt's executive order to intern Japanese Americans by six votes to three.[14]

Constitutions, like incentive systems, can be *gamed*, as some examples will show.

Article III.1 of the US constitution states that the "Judges, both of the supreme and inferior Courts, shall hold their Offices during good Behavior, and shall, at stated Times, receive for their Services, a Compensation, which shall not be diminished during their Continuance in Office." At the Federal Convention, Madison insisted that the salaries of judges should neither be diminished *nor increased* during their time in office, since Congress could bribe judges as well as threaten them. He overlooked the fact that even with the clause as it stands, Congress can game it by increasing the salaries of some judges less than those of others. In 1964, a Congress hostile to the Warren Court increased the salaries of lower federal judges by $7,500, but those of Supreme Curt Justices by $4,500 only. They were not increased again before Warren's departure in 1969.

Until the adoption of the Seventeenth Amendment in 1912, elections to the US Senate were indirect. The people of the several states elected the state legislators, who elected the US senators. Yet, as the British legal scholar James Bryce explains,

> In 1904 Oregon provided, by a law passed by the people under the initiative method of legislation contained in the constitution of that state, that the political parties might in the party primaries nominate persons for election as United States senators, and that the people might at the ensuing election of the state legislature select by their votes one of these nominees as their choice for senator. Along with this it was also enacted that a candidate for the state legislature might on his nomination either: (1) declare that he would, if elected, vote for that person as United States senator who had received the largest popular vote and thus become "the people's choice"; or, (2) declare that he would consider the popular vote as merely "a recommendation." Or he might make no

[14] In 2001 the Patriot Act passed 98 votes to 1 in the Senate and 357 to 66 in the House of Representatives.

declaration at all. In 1908 a majority of the members elected to the legislature, having made the former declaration, felt bound to carry it out, and the person who had received the highest popular vote was accordingly elected by that majority, although he was a Democrat and they were Republicans. Thus the people got their way and the federal Constitution was not formally transgressed.

The French constitution of 1958 states that "Members' right to vote [in the National Assembly and the Senate] shall be exercised in person. An Institutional Act may, in exceptional cases, authorize voting by proxy. In that event, no member shall be given more than one proxy." Since the large majority of deputies and senators have one or more local elective offices in addition to the national one (the *cumul des mandats*), they were motivated to oppose this restriction and successfully gamed it virtually from the outset. They did so by adding to the initial and precisely formulated exceptions the diffuse exception of "force majeure." Since it was left to the *bureaux* of the assembly to decide whether this case obtained, and most representatives had a strong interest in being able to absent themselves from time to time, permissions were readily granted. When consulted in 1961, the Constitutional Council naively or cynically affirmed that the *bureaux* could be trusted to verify that the constitution would be "strictly applied." In practice, the exceptions have proven far more numerous than the rule. Like the example from Oregon, this case confirms Bryce's statement that it is hard "to keep even a written and rigid constitution from bending and warping under the actual forces of politics."

I have discussed how constitutions can fail to produce the effects that their framers wanted them to have. One may also consider the opposite case, the unintended consequences of constitutions. I shall discuss three cases, which turn on the fact that the conjunction of two innocuous articles may have absurd or undesirable consequences.

According to Article I.3.4 of the American constitution, "The Vice President of the United States shall be President of the Senate." According to Article I.3.6, the Senate "shall have the sole power to try all impeachments." It follows that if the Vice President were to be impeached, he would preside over his own trial. This situation has not arisen, although it was perhaps not far from happening in the case of Spiro Agnew. According to the French constitution, on leaving office a President becomes automatically member for life of the Constitutional Council. Since nothing prevents a former President from standing for reelection, as Nikolas Sarkozy has announced he will do, and as the constitution does not require him to stand down from the Council if he were to be reelected, he might in theory be in the anomalous position of being head of the executive and also a member of the highest judicial body.

These are theoretical dangers. A vastly more consequential conjunction of two clauses that taken separately seem innocent enough is found in the Weimar

constitution. Article 48 said that "In case public safety is seriously threatened or disturbed, the Reich President may take the measures necessary to reestablish law and order, if necessary using armed force . . . The Reich President has to inform Reichstag immediately about all measures undertaken which are based on paragraphs 1 and 2 of this article. The measures have to be suspended immediately if Reichstag demands so." The last clause was presumably intended to provide a check on presidential discretion. It was undermined, however, by Article 25: "The Reich president has the right to dissolve the Reichstag, but only once for the same reason. New elections, at the latest, are held 60 days after the dissolution." Using Article 25, the President could (and did) threaten to dissolve the Reichstag should it vote to annul any measures taken under Article 48. This mechanism opened the way to power for Hitler.

In a more paradoxical case, a constitution that was not intended to "interfere with the actions of the government" suddenly became causally efficacious after a regime change, acquiring, as it were, life only after death. In Czechoslovakia, the 1968 constitution introduced, for the first time in the history of the country, a federal structure with separate assemblies (National Councils) for the Czech and Slovak lands and with strong power for these republics in the bicameral Federal Assembly. Like all Communist constitutions, this one remained a dead letter; the National Councils were not even convened. The post-Communist constitutional debates after 1990 were framed, however, by this document. The strong Slovak autonomy now became a major obstacle to reform. An amendment to the constitution required a three-fifths majority both in the proportionally elected lower house and in each of the two Czech and Slovak sections of the upper house. Although the Czech population outnumbered the Slovaks two to one, each section had 75 seats. The power of thirty-one Slovak deputies in the upper house – representing one-fifth of the voters – to block change was arguably a main cause of the break-up of the Federation in 1992. President Havel could have used his enormous moral authority to elect a unicameral constituent assembly, but he maintained the existing assembly, purged of many Communists but retaining its bicameral structure.

While constitutions may fail to have causal efficacy and may have causal effects that were not intended, these facts should not obscure the real and predictable importance of many constitutional provisions. When constitutions guarantee the overrepresentation of certain regions in parliament, as many do, one can expect these to get more roads and bridges than they would otherwise have received. If they forbid the government from instructing the Central Bank in matters of monetary policy, inflation is likely to be lower than it would otherwise have been. Unemployment, though, may be higher. If they ensure freedom of the press and free elections, the government may be reluctant to misbehave in ways that, when exposed by the media, would make it lose the

next election.[15] If they authorize patents and ban retroactive laws, there will be more economic growth. Civil and political rights create a hard core of freedoms that no constitutional jurisprudence can undo, although much will remain uncertain and sometimes arbitrary at the margins. Constitutions matter, but less than many constitutional designers and scholars believe or would have us believe.

### Bibliographical note

The problem of aligning individual and organizational incentives is the topic of J.-J. Laffont and J. Tirole, *A Theory of Incentives in Procurement and Regulation* (Cambridge, MA: MIT Press, 1994). A comprehensive handbook on corruption is A. Heidenheimer, M. Johnston, and V. LeVine (eds.), *Political Corruption* (New Brunswick, NJ: Transaction Publishers, 1989). The references to nineteenth-century English cooperatives are taken from B. Jones, *Cooperative Production* (Oxford University Press, 1894; New York: Kelley, 1968). The comment on the high turnover in government is from Chapter 22 of A. King and I. Crewe, *The Blunders of our Governments* (London: Oneworld Publications, 2013). The relative importance of trust and incentives in firms is discussed in E. Fehr and A. Falk, "Psychological foundation of incentives," *European Economic Review* 46 (2002), 687–724. The study of "earners" and "maintainers" is T. Docan, "Positive and negative incentives in the classroom," *Journal of Scholarship of Teaching and Learning* 6 (2006), 21–40. The incentive example featuring a mayor and her police chief is taken, with slight modifications, from G. Miller and A. Whitford, "The principal's moral hazard: constraints on the use of incentives in hierarchy," *Journal of Public Administration Research and Theory* 17 (2007), 213–33. Two useful studies of gaming of incentive systems are Z. Radnor, "Muddled, massaging, manoeuvring or manipulated? A typology of organisational gaming," *International Journal of Productivity and Performance Management* 57 (2008), 316–28, and D. Pitches, A. Burls, and A. Fry-Smith, "Snakes, ladders, and spin," *British Medical Journal* 327 (2003), 1436–9. The Italian norm against excellence is illustrated and explained in D. Gambetta and G. Origgi, "The LL game: the curious preference for low quality and its norms," *Politics,*

[15] The government can, however, *game* the constitutional guarantee of freedom of the press by rationing paper or printer's ink and allocating them preferentially to newspapers supporting it. In 1793, the British Attorney General tried to limit the impact of an attack on Edmund Burke by telling the author to publish his work in an expensive edition, "so as to confine it to that class of readers who may consider it coolly"; otherwise, it would be his duty to prosecute. The government can also game free elections, by choosing times and places for voting that are inconvenient to groups of voters who are likely to vote against them. I mention an example in Chapter 10.

*Philosophy & Economics* 12 (2013), 3–23. The French equivalent is exposed in F. Tagliatesta (a pseudonym for Pascal Engel), *Instructions aux académiques* (Rouen: Christophe Chomant, 2005). The study of the effects of pay-for-performance schemes is I. Siva, "Using the lessons of behavioral economics to design more effective pay-for-performance programs," *American Journal of Managed Care* 16 (2010), 497–503. The discussion of constitutions and constitution making draws on Chapter 4 of my *Securities Against Misrule* (Cambridge University Press, 2013). For the number of swing votes in Supreme Court decisions, see C. Sunstein, "Unanimity and disagreement in the Supreme Court" (unpublished manuscript, 2014).

# Conclusion: is social science possible?

## Obscurantism

Drawing on previous chapters and sometimes adding to them, I shall present my ideas about how social science should be done, and how it should not be done. Beginning with the latter, I shall criticize *soft and hard obscurantism* in the social sciences and, more briefly, soft obscurantism in the humanities. The criticism is twofold. On the one hand, obscurantism causes a massive *waste*, when students, scholars, and professionals spend years and careers learning, teaching, and practicing nonsense when they could have devoted their lives to work that would have been more useful to society and more fulfilling to themselves.[1] On the other hand, obscurantism can do *harm*, by creating the intellectual premises for actions that cause suffering or economic loss. In addition to criticizing obscurantist theories on these grounds, I shall also try, more tentatively, to explain their emergence and persistence.

As a preview of the argument, let me fill in the cells that are created if we cross the two distinctions with each other (see Table C1). I shall discuss these theories and their effects, selectively and for the most part briefly.

To document waste, intellectual criticism is sufficient. I feel confident in my objections to soft obscurantist theories and to some forms of hard obscurantism. Specifically, I know enough about science-fiction economics and political science – based on the assumption of ideally rational agents that have never existed and will never exist – to criticize them, which I shall do at some length. My objections to regression analyses are second-hand, and based on criticisms by more competent scholars. Although the appeal to authorities rather than argument is commonly, usually for good reasons, viewed as an academic sin, the stakes are so high that I am willing to risk my reputation.

[1] See a Letter to the Editor in *The Economist* for October 11–17, 2008: "Imagine what these young people [who were lured into the banking industry] could have done if they had chosen careers in science and medicine."

Table C1

| | HARD OBSCURANTISM | SOFT OBSCURANTISM |
|---|---|---|
| WASTE | Science-fiction economics<br>Science-fiction political science<br>Many regression analyses<br>Agent-based models<br>Evolutionary models | Functionalist explanation<br>Structuralism<br>Psychoanalysis<br>Analogical thinking<br>Marxism |
| HARM | The Vietnam War<br>Long-term capital management<br>The 2007 financial crisis<br>Statistical arguments to justify the death penalty and handguns | Psychoanalysis<br>Anti-psychiatry<br>Marxism |

To document harm, causal analysis is needed. I shall make some gestures in that direction, without fully substantiating my claims. Some harms are easily documented, whereas others are more uncertain. It is not clear, for instance, whether the belief in science-fiction economic theories was causally responsible for the recent financial crisis, or whether Marxism as an intellectual doctrine was causally responsible for the horrors committed in its name. I shall not try at all to justify my skepticism about agent-based models and evolutionary models, except to remark that the former seem to be getting out of hand as they are becoming increasingly opaque and that the latter have little empirical relevance.

Fighting a two-front war, as I do here, is difficult. Tocqueville wrote that the "Constituent Assembly of '89 was dispatched to fight aristocracy and despotism, and it was quite vigorous in opposing those enemies but [not] in opposing anarchy, which it was not prepared to combat ... It is rare for a man and almost impossible for an assembly to have the ability to alternately make violent efforts in two opposite directions. The energy that launched it violently in one direction impeded its progress in the other." In my case, the energy I spent over many years attacking soft obscurantism probably impeded my progress in the opposite direction. The French Revolution also illustrates another phenomenon: moderates suffer a constant risk of one group of extremists accusing them of being in league with the other, or of

each group trying to engage them as an ally in the fight against the other. These, too, are experiences in which I recognize myself.

### Soft obscurantism

A minor philosophical subdiscipline, *bullshittology*, studies the academic writings that constitute soft obscurantism. In my opinion, the study of bullshit ought to have its main location within cognitive psychology and the sociology of science, not within philosophy. Although conceptual analysis is important, the more urgent task is to *document and explain* the alarming rise of nonsense masquerading as scholarship.

Among the soft obscurantists some aim at truth, but do not respect the norms for arriving at truth, such as focusing on causality, acting as devil's advocate, and generating falsifiable hypotheses. Others do not aim at truth, and often scorn the very idea that there is such a thing. They would endorse the response of Humpty Dumpty to Alice when she said, "the question is whether you can make words mean so many different things." "The question is," he answered, "which is to be the master; that's all." *Power*, not truth, determines which theories will succeed. By assumption, these non-respecters of truth cannot be reached by argument, only by ridicule. Alan Sokal achieved this most effectively by getting an obscurantist journal to publish an article he submitted on the hermeneutics of quantum gravity, chockfull of meaningless but impressive-sounding jargon. However, this kind of hoax can work only once. I did not include these extreme obscurantists in Table C1, and shall usually ignore them in what follows.[2]

In Chapter 9, I argued that many scholars – like many of the subjects they are studying – are driven by an almost obsessional search for order and meaning in the social universe. As Albert Hirschman once remarked, they have a *seamless* vision of society, with no room for accidental benefits, coincidences, and innocent mistakes. In previous chapters I have cited two examples of such overinterpretation of behavior: the analysis of the prison system as "oppression without oppressors" (Chapter 9) and the explanation of elite norms by their efficacy in keeping outsiders out and upstarts down (Chapter 21). To elaborate on the last example, suppose it is true, as it may well be, that when intellectuals play around with language, violating rules of grammar or of spelling, they frustrate the efforts of would-be intellectuals who think they can gain access to the elite by *following* rules. From these

[2] Let me mention who they are, by discipline and by name. Disciplines include deconstructionism, postmodernism, subaltern theory, postcolonial theory, queer theory, gender theory. Some names are Jacques Derrida, Bruno Latour, Gayatri Spivak, Alain Badiou, Slavoj Žižek, Homi K. Bhabha, Judith Butler.

observations one cannot infer, however, that the playful attitude of intellectuals is *explained* by the effect on their imitators. In earlier chapters, I have repeatedly denounced such functional explanations. In Chapter 9, I also cited possible explanations of the tendency to indulge in them. Although one cannot refute an explanation by explaining why it was proposed, once it has been refuted on intellectual grounds it is legitimate to ask why it was put forward, especially if it is an instance of a larger class. I shall return to this issue of "explanations of explanations" shortly.

Concerning the humanities, I cited (Chapter 16) textual interpretations that rely on the arbitrary impression a text makes on a reader to infer the intentions of the author. I shall add an example from a famous structuralist interpretation of Baudelaire's sonnet "Les Chats." In the sixth line of the poem, we read that the "cats" of the title "cherchent le silence et l'horreur des ténèbres" (seek out the silence and the horror of darkness). The authors affirm that "not only does the word 'cat' not reappear in the poem, but even the initial fricative ["ch"] reappears only in one word ["cherchent"]. It designates, with doubling, the first action of the felines. Later in the poem, this unvoiced fricative is carefully avoided." How do the authors know that the *absence* is a deliberate *avoidance*? What is the significance of the absence? What about all the other absences one could list? What is the significance of the "doubling" of the fricative?

One of the authors of this study was Claude Lévi-Strauss, also famous for his structural analysis of myth and his invention of the idea of "mythemes," elementary components that can be combined in various ways to yield myths, just as phonemes can be combined to yield words. (The other author of the study was Roman Jakobson, an eminent phonologist.) After first applying this idea to the Oedipus myth, Lévi-Strauss later wrote four volumes, *Mythologiques*, on the myths of South American Indians. The arbitrariness of his interpretations approximates that of numerological studies that claim to retrieve "the number of the beast," 666, from the names of world leaders who would thereby be revealed to be the Antichrist. (During World War II, Hitler and Churchill were both identified as the beast.) When I had the occasion to ask one of the later occupants of Lévi-Strauss's chair whether he had students who pursued this line of analysis, he replied, "No, only he could do that." Science, however, requires intersubjectivity and replicability.

*Psychoanalysis* also lacks scientific status, in part for the same reason. An editorial in *Nature* from 2009, entitled "Psychology: A reality check," assessed it in the following terms: "Anyone reading Sigmund Freud's original work might well be seduced by the beauty of his prose, the elegance of his arguments and the acuity of his intuition. But those with a grounding in science will also be shocked by the abandon with which he elaborates his theories on the

basis of essentially no empirical evidence. This is one of the reasons why Freudian-style psychoanalysis has long since fallen out of fashion: its huge expense – treatments can stretch over years – is not balanced by evidence of efficacy." Although Freud made some valuable contributions to our understanding of the human mind (see Chapter 4), much of his work and that of his followers is vulnerable to such objections. The problem is not only lack of empirical evidence, but also lack of conceptual agreement, for instance regarding defense mechanisms. One author "reviewed 12 psychoanalytic authors, who among themselves had described 27 distinct mechanisms of defense, only 7 of which were noted by 11 of the 12 writers. [Another author] reviewed 17 psychoanalytically informed authors . . . and identified 37 different terms for defense mechanisms. Only 5 of these 37 mechanisms . . . were cited by 15 of [the] 17 authors, and only 14 of his 37 terms were cited by as many as 5 out of 17 authors."

In Chapter 9 I cited reasoning by analogy as a temptation that social scientists share with other social agents. To the many brief examples given there, I shall add the explanations offered by Marx and Tocqueville of why Christianity takes different forms in different societies. Marx inconsistently asserted both that the hoarding of gold and silver was associated with Protestantism, and that it was essentially a Catholic practice. On the one hand, "the piling-up of gold and silver gained its true stimulus with the conception of it as the material representative and general form of wealth. The cult of money has its asceticism, its self-denial, its self-sacrifice – economy and frugality, contempt for mundane, temporal and fleeting pleasures; the chase after the eternal treasure. Hence the connection between English Puritanism, or also Dutch Protestantism, and money-making." On the other hand, the "monetary system is essentially Catholic, the credit system essentially Protestant. 'The Scotch hate gold.' In the form of paper the monetary existence of commodities has only a social life. It is Faith that makes blessed." Since everything is a little bit like everything else, Marx could focus either on the fact that gold and silver, unlike credit, can be hoarded, or on the fact that credit, unlike gold and silver, depends on faith.

Tocqueville formulated a general principle "Allow the human spirit to follow its bent and it will impose a uniform rule on both political society and the divine city. It will seek, if I may put it this way, to harmonize earth with Heaven." For instance, to the fragmentation of society after the fall of the Roman Empire there corresponded a fragmentation of religion: "If Divinity could not be divided, it could nevertheless be multiplied, and its agents could be magnified beyond all measure. For most Christians homage to angels and saints became an almost idolatrous cult." Elsewhere, Tocqueville observed the opposite tendency in democracies: equality "distracts attention

away from secondary agents and focuses it primarily on the sovereign master." Inconsistently, he also claims that equality favors Catholicism, which is precisely the religion that multiplies secondary beings. As in the writings of Marx, these efforts to demonstrate an intrinsic connection between social structure and religious dogma are arbitrary. There are so many different ways of harmonizing heaven and earth and, in choosing a religion, so many more important reasons than the desire for harmony, that it is more plausible to think that the harmony comes after the event, to consolidate a choice that has already been made or imposed on other grounds.[3] As an historian of classical antiquity writes, "the belief that the unity of the Empire required monotheism by the necessity of false windows is an old sociological superstition" (see Chapter 16 on Pascal's idea of false windows).

As these examples show, Marx and Tocqueville were susceptible to the *temptation of analogies*. Unlike Tocqueville, Marx also succumbed to the *functionalist temptation*. As I mentioned in Chapter 9, Tocqueville did give in to the *temptation of agency*. These are temptations to which the human mind seems to be naturally prone. The task of the social scientist should be to resist them and explain them, not surrender to them.

As a transition to the discussion of hard obscurantism, let me observe that since functionalist explanation straddles the distinction between the soft and hard varieties, the typology in Table C1 is somewhat misleading. In Chapter 3, I cited an example of "rational-choice functionalism" to which I return later. Another example can be taken from a study of marriage and migration patterns in South India. The authors find that women tend to marry and settle down in areas that are so distant from their home region that rainfall patterns in the two regions are (somewhat) uncorrelated. This produces *risk diversification* within the extended family, since family members living in an area hit by drought can he helped out by those who are less affected. From this interesting fact, the authors conclude to the existence of "implicit inter-household contractual arrangements aimed at mitigating income risks and facilitating consumption smoothing." How an *implicit* contract can *aim* at anything, is a mystery. The authors do not seek out evidence about the

[3] Both Marx and Tocqueville also proposed explanations of *why there is religion at all*, as distinct from their attempts to explain *why there is this or that religion* in different societies. Marx said that religion was "the opium of the people," leaving it ambiguous whether this drug was invented by the dominant classes to pacify the people and prevent it from rebelling or whether the people itself created the theory of an afterlife to compensate for the miseries of this world. Tocqueville adopted the latter of these two views to explain religion in traditional societies; for democratic societies he argued (Chapter 2) that citizens need religion to compensate for the fact that they do not have a ruler. Whether correct or not, these proposals are not arbitrary in the way the analogy-based arguments are.

*explicit* reasons the women might give for their choices. Instead, they tell a functionalist just-so story.[4]

### Rational choice theory: tool-box or toy-box?

The theory of rational choice, including game theory, has immense conceptual value. In my opinion, it was the greatest breakthrough in the history of the social sciences. The idea of *maximization under constraints*, illustrated in Figure 10.1, is a simple, unifying, and powerful tool. The idea of decreasing marginal utility or marginal productivity implies that the maximization will take the form of *equalizing at the margin*. Game theory overcame a difficulty previously seen as insurmountable, by replacing the infinite regress of "I think that he thinks that I think . . ." by the concept of an *equilibrium*. In doing so, it also made intelligible why suboptimal states may persist as *bad equilibria*. In all these cases, formal modeling made it possible to convert vague preanalytical intuitions into crystal-clear understandings. I provided many illustrations in Chapters 13 and 18.

In some situations, the theory has also considerable explanatory and predictive power. Consider first explanations, using examples from Hume's *History of England*. He affirmed that the reason why some early popes created very strict regulations of divorce and marriage between relatives, up to the seventh degree of affinity, was to profit from the dispensations they could grant. As I noted in Chapter 25, he argued that the tendency of barons to stay on their estates rather than at court was individually rational behavior, although undermining the interest of their body. Finally, Hume observed that Elizabeth I, knowing that every heir would be a dangerous rival, deliberately did not name her successor. These commonsensical rational-choice explanations could obviously be multiplied indefinitely.

Consider next predictions. Rational-choice theory is an indispensable and effective tool for officials in ministries of finance and central banks when they are called upon to predict *short-term effects of small changes* in, say, tax schedules. Because consumers respond to price incentives, one can predict pretty accurately the impact on consumption of a 5 percent increase in the tax on liquor. If people equalize at the margin, they will spend part of their income on other goods. I do not think, however, that one could predict the impact of a *doubling* of the price, since the amount of smuggling and bootlegging it would

[4] The authors of this study are not first-tier economists. Yet even the unparalleled genius of Kenneth Arrow suggested that social norms are "reactions of society to compensate for market failures." Apart from the fact, which I have tried to document, that many social norms are harmful, even those that are beneficial cannot, without further argument, be explained by their benefits.

trigger depends on a host of largely unknowable factors. In the recurrent debate about legalizing hard drugs, opponents and proponents make claims about the malign or benign effects of reform that are largely unverifiable, in part because the effects would depend on how *preferences* would change if the drugs became easily available. Rational-choice theory cannot explain preferences.

Rational-choice theory turns into hard obscurantism when it ceases to be a tool-box and becomes a toy-box.[5] The examples I shall consider exhibit an uncanny *combination of mathematical sophistication, conceptual naivety, and empirical sloppiness*. Now, since substandard work can be found in any discipline, I focus on writings by economists who are highly acclaimed by their peers, recipients of either the Nobel Prize in economics or the John Bates Clark Medal. I am not claiming that all the work done by these scholars is obscurantist, only that *the publication of their obscurantist writings in leading journals or by leading publishers shows that the profession as a whole has lost its bearings.* If I had more pages at my disposal, I would follow the example of David Freedman, who reproduced, as appendices to a book on statistical models, four articles published in leading journals of economic and political science so that readers could verify whether the criticisms he made of them in the body of the text were justified. (I say more later about what he did.) As it is, I shall content myself with citing from the texts; readers are invited to seek out the original publications.

In earlier parts of this book, I have mentioned several examples of hard obscurantism:

- The claim that young people enter universities to reduce their rate of time discounting (Chapter 3).
- The claim that charitable donations and voting in national elections can be explained by the "warm glow" they produce in the agents (Chapter 5).
- The claim that social agents have well-defined and stable subjective probabilities regarding future states of the world (Chapter 6).
- The claim that the unconscious can make intertemporal trade-offs between short-term benefits and long-term costs (Chapter 7).
- The claim that tipping in restaurants can be explained as an efficient monitoring of the agents (the waiters) when it would be too costly for the principal (their employer) to do it himself (Chapter 21).
- The claim that participants in revolutionary collective action are motivated by the private benefits they will receive as leaders of the post-revolutionary society (Chapter 23).

[5] Many rational-choice models are like the steam engine invented by Hero of Alexandria in the first century AD. He considered it mainly as a toy, not as a tool that could be put to productive use. He did apparently use it, though, for opening temple doors, so unlike many of the models his engine was not completely idling.

I shall consider the first and the last of these in more detail, add another example, and then propose a general criticism.

In the model of rational choice I set out in Chapter 13, preferences are *given*, not *chosen*. They are certainly not the result of a *rational* choice, since they provide the yardstick by which action, belief formation, and information gathering can be assessed as rational. Some economists have argued, however, that people rationally choose their formal preferences (Chapter 4) – altruism, time discounting, risk attitudes – to maximize their welfare. It makes obvious sense that people's lives go better (they live longer, divorce less frequently, etc.) if they do not discount the future too heavily and are neither too risk-averse nor too risk-seeking. It is at least arguable that altruism can have the same effect: earlier, I quoted Montaigne as saying that "He who does not live a little for others hardly lives at all for himself." One might be able to characterize some formal preferences as (approximately) optimal. When an economist sees the word "optimality," however, he easily tends to read it as "rationality." Here, I consider an article arguing that the rate of time discounting is endogenous, optimal and, in fact, rationally chosen.

The basic assumption of this model is that "people have the option to put forth effort to increase their appreciation of the future" and that "more resources spent on imagination increase the propinquity of future pleasures and therefore their [present] value." For instance, a "person may spend additional time with his aging parents in order to appreciate the need for providing for his own old age." Similarly, because "schooling can communicate images of the situations and difficulties of adult life ... educated people should be more productive at reducing the remoteness of future pleasures." In fact, the authors claim that this *effect* of schooling may also provide the *motivation* to seek higher education: "more patience may be the reason why some people choose to continue their schooling." Moreover, if individuals invest in information about the afterlife there could be a spillover effect to life on earth: "To the extent that future-oriented capital is 'general' – it facilitates the imagination of events at a variety of distances into the future – a higher utility after death [*sic*] will even encourage consumption growth *before* death."

Toward the end of Chapter 6, I offered a conceptual objection to this argument: people will not invest in "future capital" unless they already care about the future. Here I shall only repeat the more elementary objection I made in Chapter 3: the neglect of the distinction between intentions and consequences. It may be true that people who spend time with their parents realize the need to provide for their old age, and that as a result they make saving decisions that make their life as a whole go better. These two causal claims provide, however, no evidence that they *intentionally choose* to spend time with the parents *in order to* learn to value the future more highly. In fact, the idea is ludicrous.

I now consider an article that aims at providing game-theoretic foundations for the transition to democracy. I shall not address all the issues discussed in the article, but only comment on the basic conceptual framework and its empirical support or lack thereof.

The authors reduce the question of class struggle to the conflict between rich and poor, thus neglecting, for instance, possible conflicts of interest between poor peasants and poor urban workers. The former have an interest in high prices on food products, the latter an interest in low prices, a phenomenon that mirrors the conflict between landowners and industrialist capitalists in nineteenth-century England. I shall not pursue this issue further, but take the two-class model as given. They also assume that all agents have identical preferences, differing only in their capital endowments. All poor agents are assumed to have the same endowments, as do all the rich. Having already swallowed the two-class assumption, why not swallow these simplifications as well? I am not equally willing, however, to accept the assumption that agents discount the future exponentially. Although mathematically convenient and seemingly justified by the hypothesis that agents are rational, the assumption has little empirical support. To adopt it without trying to justify it or defend it against criticism, which is surely well known to the authors, is a cavalier procedure.

Compared to other issues, the assumption of exponential discounting is nevertheless a minor problem. A more troublesome issue is the idea that in any given period aggregative productivity A is modeled by assuming that A is either high with probability (1-s) or low with probability s. I shall ignore the starkly dichotomous character of the assumption and focus instead on its interpretation. When the authors assert that the level of income is "stochastic," I assume they use the term in the dictionary sense of a process involving the operation of chance, such as the onset of cancer. Although scientists may today be able to quantify the probability that a given person will develop a given kind of cancer in a given period, the person herself may not – and a hundred years ago certainly could not – have any idea about the magnitude of the risk. By contrast, the authors impute knowledge of the value of s *to the agents*, in order to calculate the "discounted expected net present value ... of a poor agent after the revolution but before the state A is revealed." They would presumably defend this imputation by some version of the theory of rational expectations (Chapter 6). Whatever the defense might be, the imputation is indefensible. The idea that, say, the rural poor in France in 1789 or the urban poor in 1848 attached a sharp probability to *aggregate* productivity being high or low is a piece of science-fiction.

For the game-theoretic model of revolution to get off the ground, each class – the rich and the poor – must be viewed as a *unitary actor*. The authors raise the issue of free riding, but claim that "Because a revolution generates

private benefits for a poor agent, there is no collective action problem." In a footnote they add that "Although a revolution that changes the political system might seem to have public-good-like features, the existing empirical literature substantiates the assumption that revolutionary leaders concentrate on providing private goods to potential revolutionaries (see Gordon Tullock 1971)." The reference to the article by Tullock is strange, since it *does not offer or cite any empirical evidence* concerning actual revolutions. Tullock merely asserts that his "*impression* is that [revolutionaries] generally expect to have a good position in the new state which is to be established by the revolution. Further, my *impression* is that the leaders of revolutions continuously encourage their followers in such views" (italics mine). To cite this armchair speculation, written thirty years earlier, as a decisive piece of "empirical literature" is to offer very weak support – in fact, no support at all. Instead, the authors should have cited primary empirical sources. The French peasants who triggered the abolition of feudalism on August 4, 1789 by burning the castles of nobles in the second half of July were not motivated by the desire to occupy leading positions, nor were the East Germans who mobilized in the streets of Leipzig in October 1989 or the Egyptians who assembled in Tahrir Square in January 2011. No doubt some revolutionaries are opportunists, but it is blindingly obvious that many take risks that cannot be justified by any self-interested calculus.

Finally, I shall cite an instance of hard obscurantism from a highly regarded textbook on game theory, where the authors discuss the scope for mixed strategies (see Chapter 18). Citing the "Kitty Genovese" case, they argue that one may try to justify the idea of mixed strategies by appealing to a causal mechanism: "mixed strategies are quite appealing in this context. The people are isolated, and each is trying to guess what others will do. Each is thinking, Perhaps I should call the police . . . but maybe someone else does . . . but what if they don't? Each breaks off this process at some point and does the last thing that he thought of in this chain, but we have no good way of predicting what that last thing is. A mixed strategy carries the flavor of this idea of a chain of guesswork being broken by a random point." So far, so good.

The authors then go on, however, to commit a simple fallacy: from the correct premise that for every person there is a probability p that he will not act, they reach the false conclusion that there is a p such that each person will abstain from acting with the *same* probability p. Moreover – a second unjustified step – they assume that when all abstain from acting with probability p, their choices form an *equilibrium*, that is, that for each person the best response to all others calling the police with probability p is to call the police with probability p. The model has one seemingly attractive feature: it predicts a striking and counterintuitive stylized fact to which I have referred several times, that when the number of bystanders goes up, the probability that at

least one of them will intervene goes down. Specifically, "increasing [the size of the group] from 2 to infinity leads to an increase in the probability that not even one person helps from 0.64 to 0.8." Yet a correct prediction from absurd assumptions does not remove the absurdity. The alleged explanation is only a just-so story.

On the basis of these examples and others that I have cited throughout the book, I shall propose a selective catalogue of some characteristic procedures of hard obscurantism:

- Citing empirical evidence in a cavalier way, in the form of anecdotes, invented stories, "impressions," and unsubstantiated historical claims.
- Adopting huge simplifications that make the empirical relevance of the results essentially nil.
- Imputing to social agents mental *mechanisms* that they demonstrably do not have (exponential discounting, rational expectations, Bayesian updating), or mental *states* that they could not possibly have (well-defined subjective probabilities or complete preference orderings).
- Imputing to social agents mental *capacities* that they demonstrably do not have, such as the ability to carry out, *in real time*, calculations that take up many pages in mathematical appendices and that economists spend years mastering.
- Imputing to the unconscious mind capacities that belong only to the conscious mind, such as the capacity to weigh present and future costs and benefits against each other.
- Imputing intentions on the basis of observed outcomes.
- Imputing rationality on the basis of observed outcomes (rational-choice functionalism).
- Assuming that agents choose optimal beliefs, as assessed by the consequences of having them rather than on the basis of the evidence supporting them.
- Assuming that agents rationally choose their preferences.
- Presenting irrational behavior as rational.
- Presenting disinterested behavior as self-interested.
- Adhering to the instrumental Chicago-style philosophy of explanation, which emphasizes as-if rationality and denies that the realism of assumptions is a relevant issue.

The last procedure is the most general and probably the most important one. In Chapter 11 I cited and criticized Milton Friedman's analogy-based arguments for as-if rationality, and found them wanting. In Chapter 1 and again in Chapter 11, I surveyed claims that the non-intentional mechanisms of reinforcement and selection can mimic rationality, and found them wanting too. A defender of as-if rationality has to address the fact that the models the

mechanisms are supposed to simulate are *extremely precise and fine-grained*. The claim that market competition by and large tends to drive non-maximizing firms out of business may or may not be true, but even if true it could not support the models that fill up the pages of economic journals. Fifty thousand monkeys hitting typewriters at random over a million years might produce Scene 1 of Act I in one play by Shakespeare, but hardly the whole corpus.[6]

### Regression analysis

In this section, I often refer to the work of the late David Freedman, sometimes characterized as "the conscience of statistics" because of his relentless criticisms of facile and mechanical uses of regression analysis. Although I do not have the scholarly competence to assess his criticism – if I possessed it, I would not need to use him as a crutch – it corresponds to much I have observed in the course of a long academic career.

In an influential article on "Statistical models and shoe leather," Freedman states his general position as follows:

> A crude four-point scale may be useful: 1. Regression usually works, although it is (like anything else) imperfect and may sometimes go wrong. 2. Regression sometimes works in the hands of skillful practitioners, but it isn't suitable for routine use. 3. Regression might work, but it hasn't yet. 4. Regression can't work. Textbooks, courtroom testimony, and newspaper interviews seem to put regression into category 1. Category 4 seems too pessimistic. My own view is bracketed by categories 2 and 3, although good examples are quite hard to find.[7]

Regression analysis has an almost infinite number of potential temptations, pitfalls, and fallacies. Let me cite a few: data mining (shopping around for independent variables until one gets a good fit), curve-fitting (shopping around for a functional form that yields a good fit), arbitrariness in the measurement of independent or dependent variables, sample heterogeneity, the exclusion or inclusion of "outliers," selection biases, the use of lagged variables, the problem of distinguishing correlation from causation, and that of identifying the direction of causation.[8] In addition, very importantly, to ensure the *quality*

[6] Needless to say, this is only a rhetorical statement. However, an article examining whether "a consumer might be able to find a reasonably good 'rule-of-thumb' approximation to optimal behavior by trial-and-error methods as Friedman . . . proposed long ago" found that "individual learning methods can reliably identify reasonable search rules only if the consumer is able to spend absurdly large amounts of time searching for a good rule."

[7] When I have presented my objections to data analysis to various audiences, my critics have usually located themselves at point (1) of this scale.

[8] In what may be the earliest sustained criticism of statistical modeling, Keynes criticized the Dutch economist Tinbergen for "fidget[ing] about until he finds a time-lag which does not fit in too badly with the theory he is testing" and also for (what is now called) curve-fitting. In

of the data, scholars have to engage in demanding "shoe leather" work that they may – consciously or unconsciously – resist. These problems – of which I have cited only some of the best known – are too numerous and varied to be fully covered by a textbook exposition, even at an advanced level. There are certain general lessons, such as testing for "robustness," but the number and variety of tests to run is a matter of judgment and experience. Scholars simply have to learn by trial and error until they know what tends to work. Regression analysis is not a science, nor – as is sometimes asserted – an art, but a *craft*. It is guided by informal norms shared by elite scholars rather than by formal rules that can be mechanically applied. To learn the craft properly, a practitioner has to work through hundreds, perhaps thousands, of applications. The process of testing and eliminating counterhypotheses is a subtle skill that cannot be reduced to rote.

More importantly, perhaps, for all but exceptionally gifted scholars, mastering the craft is so time-consuming and demanding that it excludes the acquisition of substantive knowledge in any broad field of empirical inquiry.[9] At the same time, substantive knowledge is often indispensable. Among the various problems I enumerated above, the crucial one of distinguishing causal from spurious correlations may require deep familiarity with the field in question, in order to know which among the indefinitely many possible variables one should include as controls in the regression equations. To take a simple example, a person unfamiliar with geometry might try to estimate the area of rectangles as a function of their perimeter. Drawing twenty typical rectangles and doing the regression, he finds a correlation coefficient of 0.98. In a similar example, he might try to estimate the surface area of randomly selected cylinders and cones as a function of their radius and height, and find a significant relationship. In both cases the correlations would be spurious and non-predictive. In these examples, to be sure, the correct understanding is a matter of logic, not of causality. They serve only to illustrate the point that in the absence of substantive knowledge – whether mathematical or causal – the mechanical search for correlations can produce nonsense.

I conjecture that a non-negligible part of empirical social science consists of half-understood statistical theory applied to half-assimilated empirical material. To substantiate this assertion, I refer to David Freedman's detailed analyses of six articles published in leading academic journals: four from *American*

addition, as I have mentioned, he criticized the assumption that agents maximize expected utility. In other words, Keynes objected to both forms of hard obscurantism that I discuss here.

[9] An argument in a letter by the biologist Stuart Firestein to *The Economist* (November 9, 2013) may apply even more strongly to the social sciences than to his field: "Demanding that scientists be sophisticated statisticians is as silly as demanding that statisticians be competent molecular biologists or electrophysiologists. Both are professional abilities that are not likely to be mastered by the same people."

*Political Science Review*, one from the *Quarterly Journal of Economics*, and one from *American Sociological Review*. The number of mistakes and confusions that he finds – some of them so elementary that even I could understand them – is staggering. It would be tempting to dismiss his criticism by responding that "substandard work exists everywhere." Yet, commenting on three of the articles, Freedman writes that they "may not be the best of their kind, but they are far from the worst. Indeed, one was later awarded a prize for the best article published in *American Political Science Review* in 1988." If a substandard article can not only pass peer review in the leading journal of the profession but also be deemed "best of the year," one must wonder, as I did earlier, whether the profession has lost its bearings.

Next, I refer to Freedman's comments on how to avoid data mining. From my limited experience I have concluded that even when scholars try to be honest and not rig the cards in their favor, they may unconsciously favor definitions and measurements that favor the hypothesis they want to be true.[10] To keep this tendency in check, the scholar could use either replication or cross validation ("out-of-sample testing").[11] The former, according to Freedman, is "commonplace in the physical and health sciences, rare in the social sciences." The latter takes the following form: "you put half the data in cold storage, and look at it only after deciding which models to fit. This isn't as good as real replication but it's much better than nothing. Cross-validation is standard in some fields, not in others." As far as I can gather, this method is not standard in the applied social sciences. It is not recommended in textbooks or required by journal editors. An alternative form of self-binding – probably too utopian to be seriously considered – would be to post the hypothesis to be tested on the internet ahead of testing it.

Freedman's rigorism did not please everybody. Let me cite his hilarious caricature – which like any good caricature reveals important features of its object – of the responses that modelers made to his criticism:

**The modelers' responses**

We know all that. Nothing is perfect. Linearity has to be a good first approximation. Log linearity has to be a good first approximation. The assumptions are reasonable. The assumptions don't matter. The assumptions are conservative. You can't prove the

[10] The problem can also arise without any individual bias. As two scholars note, "it is not necessary for any one researcher to mine the data deliberately. It suffices that several researchers independently consider alternative predictors and only significant results are published."

[11] The same scholars note, however, that "Nothing ensures that the researcher who presents pseudo out-of-sample validation results in his paper has not experimented with other predictors without showing the results." My conjecture is that this procedure would require conscious rather than unconscious manipulation, and hence is less likely to occur. Even if this conjecture is correct, it would not address the issue of collective data mining presented in the previous note. However, my main focus here is on the problems, not on the efficacy of remedies.

assumptions are wrong. The biases will cancel. We can model the biases. We're only doing what everybody else does. Now we use more sophisticated techniques. If we don't do it, someone else will. What would you do? The decision maker has to be better off with us than without us. We all have mental models, not using a model is still a model. The models aren't totally useless. You have to do the best you can with the data. You have to make assumptions in order to make progress. You have to give the models the benefit of the doubt. Where's the harm?

I return to the question of harm shortly. First, however, I conclude the present section by discussing two procedures that are supposed to reduce the scope for arbitrariness and subjectivity in statistical analysis: randomization and the use of "instrumental variables."

In applied policy analysis, we seek to answer, not "What worked in the past," but "What will work?" The former question can be addressed by regression analysis, for instance by looking, within a given country, at communities with different rates of child mortality (dependent variable) and try to identify the causes (independent variables). The latter question could be addressed by first conjecturing that child mortality can be reduced (say) by providing free mosquito nets, then choosing at random, within a given country, one set of communities that will receive the nets (the treatment group) and another that will not (the control group), and finally observing whether the treatment group has significantly lower child mortality than the control. (It is not obvious that it will have, since people may not value goods they get for free.) If it does, we can conclude not only that the provision of nets should be generalized to all communities, but also that it *explains* the lower mortality, since the randomization effectively excludes other causes. From the policy point of view, a limitation of this approach is that the recommendation to provide the nets cannot be generalized to other countries. From the explanatory point of view, a limitation of the approach is that it tells us nothing about *how* the treatment affected the outcome. The causal explanation is a black box (Chapter 2). The two limitations are related, since if we understood the causal mechanism producing the treatment effect, we would be better placed to assess its usefulness in other countries. A more general limitation is obviously that the method cannot be used to explain events *in the past*.

The use of instrumental variables is supposed to overcome the last limitation while also overcoming the limitations of standard regression analysis. Roughly speaking, the procedure relies on *natural experiments*. Consider, for example, the question whether, for a given crime, judges or juries are more likely to acquit the accused in criminal trials. In countries where both procedures are used, one might try to settle the question by simple regression analysis. It is possible, however, that judges or juries are unequally influenced by other factors, such as the age, sex, race, or physical appearance of the accused; the existence of mitigating or aggravating factors that in theory should influence

only the sentencing, but may also shape the verdict; the sentence he or she would receive if found guilty; and so on. I do not think one could gather the data needed to control for such factors, and in any case there could be other factors. In France, a natural experiment from 1941 proved pretty conclusively that judges were more severe than juries. In that year, the Vichy government reduced the number of jurors from twelve to six, and supplemented them with three professional magistrates. The acquittal rate fell from 24.7 percent in 1941 to 8.4 percent in 1942. It is highly plausible that the new law was the cause of the fall.[12]

Using instrumental variables, recent scholarship has provided many ingenious and plausible demonstrations of causality. This research has an unfortunate bias, however, since scholars tend to be attracted by cases where a *natural experiment happened to occur* rather than by cases with intrinsic intellectual or social importance. One can apply to natural experiments a phrase that I used in Chapter 24 about performance targets: they are good servants, but bad masters. When no servant can be found, scholars have to use their own shoe leather.

### Waste and harm

I said earlier that obscurantism is capable of causing both waste and harm. The distinction may seem specious, since waste on a large scale has opportunity costs, by preventing the prevention of harm. Be that as it may, I understand *waste* as the social cost of the *teaching* and *practicing* of obscurantist theories. These may be defined either as direct costs (salaries of teachers and practitioners, tuition expenses) or as opportunity costs (the contributions that teachers, practitioners, and students could have made in other activities). Here, I use the first and more tangible definition. I mostly define *harm* as the avoidable suffering or loss that professionals who give advice based on obscurantist theories impose on others. I shall also, in a more speculative vein, discuss whether the proponents of political theories, such as Marxism, can be said to have caused harm.

In principle, it is not impossible to quantify the social waste caused by the teaching of *soft* obscurantism. As a minimum, one could estimate the number of scholars whose research is based on soft obscurantist theories and multiply it by the average salary of a college professor. Some years ago, I got into trouble in Norwegian media when I made a back-of-the envelope calculation of the social costs of soft obscurantism in Norway, and proposed the number of $15 million per year (200 teachers). A similar calculation for hard

[12] A reduction of the acquittal rate was also the principal motive behind the reform. Justifying it, Minister of Justice Joseph Barthélemy said that "although it does not suppress the jury, it tends to defang it."

obscurantists would be more difficult, since many of its practitioners also do useful work. Unlike soft obscurantists, they have *skills,* which may be put to good as well as bad uses. A practitioner of science-fiction economics may also design auction systems that save huge sums for society.

The question is somewhat trivial. One might also want to deplore the waste created by many forms of modern art, including the employment of curators who think they must add something to the works of art rather than taking care of them.[13] Other deplorable activities include the massive advertising directed at young girls in contemporary societies to make them attach an irrational importance to physical appearance. (In their case, the effect is not only waste of money, but also harm done to those who do not live up to the standards.) No society known to me has been exempt from frivolous, fashionable, expensive, and pointless activities of this kind. To complain about them is a bit like complaining about the weather.

*Harm* is a much more serious issue. Doctors have always been taught *primum non nocere* – above all do no harm. Although it has been said that up to the mid-nineteenth century most medicine was *iatrogenic* – causing illness by treatment intended to alleviate it – these practices could to some extent be excused by ignorance. The excuse is incomplete, because in many cases the doctors should have known that they were ignorant. Writing in the sixteenth century, Montaigne was perfectly aware of the lack of evidence-based medicine: "even if a cure is achieved, how can the doctor be certain that the malady had not simply run its course or that it was a chance effect or produced by something else the patient had eaten, drunk or touched that day? Furthermore, if that proof were absolutely convincing, how many times was it repeated and how often was the doctor able to string together such chance encounters again, so as to establish a rule?"

Montaigne had three *bêtes noires*: doctors, lawyers, and scholars. Because harm done by lawyers does not usually result from their reliance on obscurantist theories, I shall ignore them here. I shall discuss, though, cases in which such reliance may have caused *judges* to do harm. Concerning scholars, a dictum by Montaigne inspired the present book: "It may be plausibly asserted that there is an infant-school ignorance which precedes knowledge and another doctoral ignorance which comes after it."[14]

[13] I once visited a French building with beautiful Romanesque capitals where the curator wanted to create an atmospheric effect by dimming the lights, making it impossible to see the all-important details of the sculptures.

[14] Pascal, expanding on Montaigne, put it more eloquently: "Knowledge has two extremes which meet; one is the pure natural ignorance of every man at birth, the other is the extreme reached by great minds who run through the whole range of human knowledge only to find that they know nothing and come back to the same ignorance from which they set out, but it is a wise ignorance which knows itself. Those who stand half-way have put their natural ignorance behind them

I now discuss some ways in which presumed knowledge has caused harm, beginning with the more straightforward cases.

*Psychoanalysis* certainly leads to a waste of time and money. The *Nature* editorial cited earlier observed that the "huge expense – treatments can stretch over years – is not balanced by evidence of efficacy." Things are actually worse, however, since there is considerable evidence of psychoanalysts or psychodynamic (broadly Freudian) therapists causing harm, mostly by misattributing causation of mental illness and of deviant behaviors. Per capita, the French probably suffered the most. In the aggregate, Americans have probably been exposed to the most harm.

Several theories, inspired by psychoanalysis, blamed "cold" or absent mothers for the problems of their children. (These theories may or may not be related to Freud's documented misogyny.) Some writers claimed that schizophrenia was caused by "schizophrenogenic" mothers, others that emotional problems in adults could be traced back to the fact that their mothers worked, rather than staying at home, when they were children,[15] and still others that the "refrigerator mother" was responsible for autism. For reasons of space, I limit myself to the last.

Autism is a neurodevelopmental disorder that affects one or two persons per thousand. In 1967, Bruno Bettelheim, an Austrian philosopher with no medical or psychological training, highly influenced by psychoanalysis, published *The Empty Fortress: Infantile Autism and the Birth of the Self*, in which he claimed that autism was caused by emotionally cold parents. In his opinion, "the precipitating factor in infantile autism is the parent's wish that his child should not exist." Although a complete charlatan, he became a professor at the University of Chicago, was elected to the American Academy of Arts and Sciences, and had an undeserved reputation as a wise and humanitarian psychologist. In the 1970s and 1980s, his views on autism were highly influential in Western societies, causing an untold number of parents to develop acute guilt feelings because they believed they were responsible for their child's illness. The blame for this harm must be laid squarely at the door of the American psychological community. If the peers who certified him had had the most elementary notions about what counts as science, he would have suffered the fate of most quacks.

without attaining the other; they have some smattering of adequate knowledge and upset everything. They upset the world and get everything wrong."

[15] This claim was central to "attachment theory," the outcome of the mutually supportive work of John Bowlby and Konrad Lorenz. Intellectually, the theory has been widely criticized as speculative. In a revealing comment, Bowlby claimed that there are "two groups with a vested interest in shooting down the theory. The Communists are one, for the obvious reason that they need women at work and thus their children must be cared for by others. The professional women are the second group. They have, in fact, neglected their families. But it's the last thing they want to admit."

Bettelheim is totally discredited in the United States, but his views on autism remain influential in France, which is, with Argentina, the main bastion of psychoanalysis today. Although the physiological and genetic roots of the syndrome had been known for decades, it was only in 2012 that the French Haute Autorité de Santé asserted that psychoanalytic treatments of autism were "not recommended." This decision followed upon a heated polemic around a film, *Le Mur*, which showed interviews with eleven French psychoanalysts, several of them close to the ultra-obscurantist schools of Jacques Lacan and Melanie Klein. The analysts came across as massively irresponsible, one of them saying that: "I don't care (*ça m'est égal*) if the child does nothing during the whole session while I keep dozing next to him, I am used to this [*sic*] in my work as a psychoanalyst." Nevertheless, upon the request of some of the analysts a lower court censured the movie, until its decision was lifted in 2014.

For many years, French psychologists and psychiatrists with a background in psychoanalysis resisted the use of methadone in the treatment of heroine addicts. A French psychiatrist, who was himself influenced by psychoanalysis before he came to his senses, writes that "In retrospect, it is clear that psychoanalysis, combined with the weight of moralizing, prejudices, ignorance and special interests, for many years prevented the establishment of an effective treatment of drug addicts. In France, more than 10,000 lives might have been saved if there had not been, for almost twenty years, this wall of resistance." Although the number of 10,000 may suffer from excessive precision, the analysis leaves no room for doubt about the harms caused by the French psychoanalytical community. The withdrawal symptoms of addicts were routinely interpreted and treated as existential anxiety. Reflecting on France's delay in methadone treatment, a colleague of the author said "it is because the French have not bothered to read English and study the American literature thinking, with Lacan, that they were the world leaders in Theory."[16]

The theory of the "repressed memory syndrome," another descendant of Freudianism, has also caused great harm. According to the scholar who has done the most to denounce this theory, many of the alleged memories are simply false, not repressed:

> The fact that the memories of victims and witnesses can be false or inaccurate even though they believe them to be true has important implications for the legal system and for those who counsel and treat victims of crimes. Some psychotherapists use techniques that are suggestive (along the lines of, "you don't remember sexual abuse, but you have the symptoms, so let's just imagine who might have done it"). These can lead patients to false beliefs and memories, causing great damage to the patients themselves

[16] If the predominantly Anglo-American readers of the present work feel that their communities are immune to such obscurantism, they might reflect on Bettelheim and ask themselves who, today, have taken on his mantle.

and to those who are accused. In one Illinois case, psychiatrist Bennett Braun was accused by his patient, Patricia Burgus, of using drugs and hypnosis to convince her that she possessed 300 personalities, ate meat loaf made of human flesh and was a high priestess in a satanic cult. By some estimates, thousands of people have been harmed in similar ways by well-meaning providers who apply a "cure" that ends up being worse than the disease.

The use of psychologists trained in a psychodynamic tradition as expert witnesses in court can also cause a great deal of harm. In a Norwegian case where a father was accused of sexual abuse of his daughter, on the basis of her statements, an expert psychologist testified that the sharp fence posts in the child's drawing of a house surrounded by a fence very likely had a sexual significance (*Aftenposten*, Oslo, October 9, 1999). She affirmed, moreover, that the number of posts in the fence very probably indicated the number of occasions on which the child had been abused. The child's father spent two weeks in jail, in a security cell, was barely acquitted of incest, but his life was ruined. Later, the daughter confessed that it was all an invention.

A further example of the possible harm done by soft obscurantism is conjectural, but I believe it would be worthwhile to explore it. I have in mind the effects of the anti-psychiatric movement of the 1970s, led or inspired by Michel Foucault, Ronald Laing, Thomas Szasz, and Franco Basaglia, among many others. Like Bettelheim, these authors were clearly obscurantists in denying the "hard" (genetic or neurological) basis of many mental illnesses. It is, I think, uncontroversial that there was some connection between this "movement" (it was not really organized) and the dismantling of large psychiatric units in several countries. Whether the movement *caused* the process or the two were effects of a common cause – such as "the spirit of the sixties" – remains to be determined. It seems also uncontroversial that some inmates benefited from the change, while some chronically ill patients suffered. Whether the net effect was positive or negative, also remains to be determined.

My final example of the effects of soft obscurantism is even more speculative, and probably not amenable to empirical investigation. I have in mind the theories of Karl Marx, as stated by him, not as developed by his successors. Marx was certainly not causally responsible for all the actions that the leaders of the Soviet Union and China committed in his name. Their references to the ever-flexible "Marxist" doctrines were mostly rationalizations for what they wanted to do anyway, essentially to hold on to power. The main exception I can think of arises from Marx's utter commitment to the idea that Communist society was the *end* of history, in both senses of that term. To the (probably unknowable) extent that Lenin, Stalin, and Mao Zedong took over this teleological conception of history, it may have made them more willing to sacrifice millions of lives for the sake of eternal Communist bliss.

The harm caused by hard obscurantism is far from easy to determine. I shall consider three cases: the impact of obscurantist social science on the conduct of the Vietnam War, the impact of obscurantist statistical analyses on decisions by the American Supreme Court, and the impact of obscurantist economic models on the recent financial crises.

The Vietnam War caused the death of 58,000 American soldiers, perhaps a million Vietnamese soldiers and civilians, and several hundred thousand civilians in Laos and Cambodia. Financial costs to the US are estimated to $700 billion. Vietnam was physically devastated. If these losses could be charged to hard obscurantism, the indictment would be devastating. The question is probably too complex to admit of a clear answer. The major mistakes in the war stemmed from false analogies, ignorance about Vietnamese nationalism, the belief that international Communism was a monolithic bloc, adherence to the intellectually flawed domino theory, wishful thinking about the South Vietnamese army, and American electoral considerations, not from hard obscurantism. Yet the obsession with quantification in decision makers and advisers may have prevented them from looking in the right place, as the proverbial drunk who looked for his lost key under the lamppost because it was the only place where there was light.

The quantification extended to facts, probabilities, and utilities. I shall cite some sample statements by Robert McNamara (Secretary of Defense), John McNaughton (his assistant), McGeorge Bundy (National Security Adviser), William Bundy (Assistant Secretary of State for the Far East) and Walt Rostow (McGeorge Bundy's successor as National Security Adviser).

"I think that without this decision [to commit troops in Vietnam] the whole program will be half-hearted. *With* this decision I believe that the odds are almost even [*sic*] that the commitment will not have to be carried out." (McGeorge Bundy, November 1961)

"Every quantitative measurement we have shows we're winning the war." (McNamara, late 1962)

"We cannot assert that a policy of sustained reprisal will succeed in changing the course of the contest in Vietnam. It may fail, and we cannot estimate the odds with any certainty – they may be somewhere between 25 percent and 75 percent. What we can say is that even if it fails, the policy will be worth it. At a minimum it will dampen down the charge that we did not do all that we could have done, and this charge will be important in many countries, including our own." (McGeorge Bundy, February 1965)

US aims were defined as follows. "70 percent – To avoid a humiliating US defeat (to our reputation as guarantor). 20 percent – To keep SVN [South Vietnam] and then adjacent territory from Chinese hands. 10 percent – To permit the people of SVN to enjoy a better, freer way of life." (McNaughton, March 1965)

If "we assume that the situation is deteriorating so that we have at present no more than a 20 percent chance of stemming the VC gains so that Hanoi would come to terms, we

believe that the introduction of major additional US forces would not increase the chances of success to more than 30 percent, and would run the overwhelming risk of a truly disastrous US defeat. We believe such defeat would be far worse than defeat without such a major additional commitment." (William Bundy, June 1965)

McNaughton thought the "chances of victory with that number of troops [200,000 to 400,000+] would be 20 percent throughout 1966, 40 percent in 1967, and still only 60 percent by election year in 1968. How did one value-scale the desirability of various outcomes? McNaughton asked himself. 'Is a collapse at a 75,000 level worse than an inconclusive situation at 200–400,000 level? Probably yes.'" (July 1965)

The "Communists were losing the battle for hearts and minds at a rate that had now reached 3 percent of the population a month." (Rostow, November 1968)

As I argued in the Introduction to Part II, the precision of such assessments is spurious. It is hard to tell, of course, whether they were used as premises for action or were rationalizations for decisions made on other grounds. In either case, they undermined the non-quantified advice offered by those who knew the history, culture, and language of Vietnam.

Turning next to the possible harm caused by statistical analysis, consider the impact of Chicago-style economics on legislation concerning the death penalty and gun control. In one summary, "Isaac Ehrlich's analysis [in 1975] of national time-series data led him to claim that each execution saved eight lives. Solicitor General Robert Bork cited Ehrlich's work to the Supreme Court a year later, and the Court, while claiming not to have relied on the empirical evidence, ended the death penalty moratorium when it upheld various capital punishment statutes in *Gregg* v. *Georgia* and related cases." Ehrlich's work was later discredited. The claim by John Lott – cited by John Ashcroft when he was Attorney General in the Bush administration – that the right to carry concealed handguns saves lives is also dubious and seems to be driven mainly by ideology. Commenting on Lott's work, one scholar writes that "The academic survival of a flawed study may not be of much consequence. But, unfortunately, the ill-effects of a bad policy, influenced by flawed research, may hurt generations." In other words, we can tolerate waste, but we should not accept harm.

Regarding financial crises, I do not know of any detailed study discussing the importance of modeler hubris in the 1998 collapse of Long Term Capital Management, costing investors $4.5 billion, or in the current financial crisis. Greed, short-termism and deregulation may have been more important than unwarranted confidence in the Nobel Prize-winning models. To be sure, some fund managers told their clients that according to their models a crisis of the magnitude of what has happened since 2007 would only occur once in 100,000 years. It remains to be shown, however, that these managers actually *believed* in the models and used them as premises for decisions or advice. After all, being "too big to fail" they had very little to lose if the models got it wrong and

much to gain if they did not (a mechanism that has been called "survival of the fattest," rather than of the fittest). They may have been dishonest rather than incompetent, crooks rather than stupid. That being said, I find it hard to believe that excessive belief in the efficiency of markets did not play some role in generating the crisis. Since the information that is reflected in asset values is a public good, nobody has an incentive to produce it. This free-rider problem has arguably generated mechanical diversification of assets as a substitute for due diligence.

### Explaining obscurantism

Several of the preceding comments on harm done by soft and hard obscurantism were conjectural – sketches of a research program rather than statements of facts. This characterization is equally apt for the present section. Both the psychological explanations I offer of the emergence of obscurantism and the sociological explanations I propose of its persistence are at best strongly suggestive, not conclusive.

In Chapter 9, I argued that scholars sometimes go wrong because of the strong tendency of all human beings to find meaning and order in the world, causing them to search for agency, objective teleology, and analogy. Before elaborating, it seems appropriate, in a chapter where I criticize other scholars, to recount some of my own failings. On three occasions in the 1970s, I fell victim to the lure of analogy, and perhaps also of teleology. In a book in Norwegian from 1971, I drew a parallel between the impossibility of predicting technical change and Gödel's incompleteness theorem. When a logician colleague at my university raised his eyebrows, I realized the foolishness of the analogy. In a book in French from 1975, I approvingly cited Jacques Lacan's analogy between Marx's concept of surplus-value (*Mehrwert)* and Freud's concept of surplus-pleasure (*Mehr von Lust).* I did not need help to quickly realize how stupid that comparison was. In a book from 1979, I claimed that the system of periodic elections without the possibility of recalling representatives "can be interpreted [as] the electorate's method of binding itself and of protecting itself against its own impulsiveness." Needless to say, no electorate ever did anything of the kind. In that case, the flaw in my reasoning may have been due either to a misplaced analogy between individual and collective self-binding or to objective teleology. In getting rid of my confusion, I was assisted by a history professor who told me bluntly, "In politics, people never try to bind themselves, only to bind others." An irony is that I proposed this "interpretation" in a book, *Ulysses and the Sirens*, which among other things was a crusade against functionalist explanations.

As I argued in Chapter 9, explanation by agency, objective teleology or analogy can produce a click in the brain that is easily confused with the click of

explanation – *The pleasure of finding things out*, as Richard Feynman called it in a book of that title.[17] The production of some kind of click may also be why people find patterns in random sequences of Heads and Tails in coin tosses. In a classic study of perceptions of randomness we read that "among the 20 possible sequences (disregarding direction and label) of six tosses of a coin, we venture that only H T T H T H appears really random. Or four tosses, there may not be any." As I mention shortly, it has been argued that the left brain hemisphere is involved in this process of finding spurious patterns.

The tendency to search for patterns is obviously only a necessary condition for obscurantism, not a sufficient one. We all have it, and we are not all obscurantists. What, in a given person, actualizes our common potential for talking and writing nonsense? Alternatively, what prevents us from doing it all the time? According to a study I cited in Chapter 9, teleological explanation seems to be the default mode. According to neuroscientists, the study of the brain might explain both excessive pattern seeking and the fact that it is kept within limits. The left hemisphere has the function of imposing a coherent framework on all the information with which we are constantly bombarded. It is corrected by the right hemisphere, which serves as devil's advocate and dismantles the constructions of the left hemisphere when they get out of bounds. I am reluctant to pursue these speculations, fascinating as they are. In the abstract, it seems plausible that natural selection has favored both a tendency to jump to conclusions and a tendency to correct the first tendency when it gets out of hand, but this piece of armchair reasoning has no purchase on the explanation of obscurantism. Too many intermediate links in the causal chain are missing.

Hard and soft obscurantism often have in common what, citing Albert Hirschman, I called their *seamless* character. In some forms of soft obscurantism, Western, capitalist, male, and heterosexual domination accounts for all social phenomena, *with no residual*. In some forms of hard obscurantism, rational choice and self-interest account for all social phenomena, *with no residual*. These two statements exaggerate somewhat, but not wildly. They suggest a possible tendency of the human mind to search for grand unifying theories and to disregard stubborn facts that do not fit, a mindset not amenable to the piece-meal mechanism approach that I advocate in this book. I cannot think, however, of an evolutionary explanation for this tendency, if it exists, or even a plausible just-so story.

Turning now from the *emergence* of obscurantism to its *persistence*, we have to confront the claim that science is a form of organized skepticism that sooner or later will weed out invalid theories. In the humanities everywhere and in the social sciences outside the Anglo-American sphere of influence, soft

[17] The book includes a lecture on "Cargo cult science," which is a good study of the psychological roots of obscurantism.

obscurantism is a strong presence and shows no sign of abating. In that sphere of influence, hard obscurantism has achieved a seemingly impregnable status in economics and political science, and to a lesser degree in sociology. In the natural sciences, the Ptolemaic system of astronomy, phlogiston theory, phrenology, the theory of spontaneous generation, and Lamarckism, were eventually weeded out. Why are the social sciences so different?

One might put one's hope in the word "eventually." After all, the teleological Aristotelian physics dominated Western thought for two millennia until someone thought of looking out of the window. Theories of alchemy, which also existed for millennia, counting Francis Bacon and Newton among their believers, made extensive use of analogies ("correspondences") until swept away by Mendelev. Alfred Wallace, the co-inventor of the theory of natural selection, believed in spiritism. Perhaps the social sciences of the twenty-fifth century will be non-obscurantist. In the meantime, though, I would like to understand why obscurantism shows no sign of fading away.

Among the theories I mentioned, phlogiston theory, phrenology, the theory of spontaneous generation, and Lamarckism were refuted by the *facts*. The Ptolemaic system crumbled under the weight of the complex constructions needed to "save the appearances," that is, to accommodate the facts, and by the proposal of a simpler alternative theory. Most theories in the social sciences stand in little danger of being refuted by the facts. This is obviously true for soft obscurantism, but no less for the hard variety. On *central* issues, competing schools of economists *disagree completely*, as shown by the absurd award of the Nobel Prize in economics in 2013 to one economist who had predicted the recent financial bubble and to another who denied that there was a bubble. In statistics, the battle between Bayesians and frequentists seems never-ending. The Nobel Prize in physics is awarded only to scientists who have made confirmed predictions that are not also consequences of rival theories, which is why neither Stephen Hawking nor the string theorist Edward Witten has received it. Many of the economists who have received the Alfred Nobel Memorial Prize for Economic Science work within the paradigms of rational choice theory and statistical modeling. Yet *not one of them was awarded the prize for confirmed empirical predictions*.[18] By an ironic contrast, on the one occasion it was awarded on that basis it went to Daniel Kahneman for his work in behavioral economics, notably for the discovery of loss aversion.[19]

[18] One economist received the prize for the theory of "the market for lemons," which was later *disconfirmed* by behavioral economists. The theory predicts that people will not buy what might be a lemon, such as a used car, but the "winner's curse" (Chapter 14) shows that they do. The economist later embraced behavioral economics.

[19] I do not think confirmed predictions should be the only criterion. *Adding to the toolbox of mechanisms* can be equally valuable. Among the Nobel Prize winners in economics Thomas Schelling is the outstanding example.

The invulnerability to empirical objections may be a necessary condition for the persistence of obscurantism, but hardly a sufficient one. I shall propose five sociological mechanisms that may add some explanatory power.

One obscurantism-sustaining mechanism is *mind binding*, an idea conceived on analogy with the Chinese practice of foot binding, which persisted as a bad equilibrium for centuries.[20] Given that no parents would let their son marry a woman who did not have her feet bound, it was in the interests of the parents of girls to adhere to the practice. Although crippling and horribly painful, the practice was sustained by the fact that no family had an incentive to deviate unilaterally. My observation of the American academic situation strongly suggests to me that departments of economics and, increasingly, political science are caught in a bad equilibrium of this kind. The mind binding to which they subject their students is due, at least in part, to the perceived need to produce marriageable – hirable – candidates. A department that reduced the course load in game theory and data analysis while increasing the course requirements in economic or political history would have difficulties placing their students in first-tier universities.[21]

A second mechanism arises through pluralistic ignorance (Chapter 22). In the case of economic and statistical models, this situation would obtain if most scholars, although secretly worried about the procedures, kept quiet because of the perception that most of their colleagues were firmly convinced of their validity. There are several mechanisms that might be at work here. From my own experience I know very well how a scholar's confidence in his own judgment can be undermined by the fact that the majority thinks differently. *How could all these people, who are certainly smarter than I am, be wrong?* Also, even with unshakable self-confidence, a scholar might worry that speaking up might cause ostracism and career obstacles. I am more concerned, however, with self-doubt than with opportunism.

A third mechanism derives from the use of citation rankings to allocate funds to universities, departments, or individual scholars. Although many writers have commented on the perverse and pathological features of this

[20] The analogy is *not* an instance of the first law of pseudo-science (Chapter 9), but reflects the fact that foot binding and mind binding have the same formal structure: no agent has an incentive to deviate unilaterally from the bad equilibrium.

[21] Earlier, I mentioned string theory as an instance of a physical theory with no confirmed predictions. One would think that a department that contained a mix of string theorists and other particle theorists would be healthier than one in which all the particle theorists subscribed to string theory. This is, for instance, the view of Gabriele Veneziano, a co-inventor of string theory. Yet the dominance of string theory persists as a bad equilibrium. For American Ph.D. candidates to be marriageable, that is, capable of being hired as particle theorists in a research university, they *must* work in string theory. The prestige of the theory is probably due to its mathematical complexity and beauty. Witten was awarded the Fields Medal, the most prestigious prize in *mathematics*. The Nobel Prize committee for *physics* rightly disregards this feature of the theory.

system, they have not, to my knowledge, noted its obscurantist-sustaining effects. Once a group of obscurantist scholars in a given field has reached a critical mass, the number of within-group citations can be used to argue for funds and positions that will further cement their grip on the discipline in question.

Fourth, obscurantism is sustained by informal social norms ("don't rock the boat") of the academic community that prevent frank criticism. In Norway, where I have criticized soft and hard obscurantism on many occasions, I am regularly accused of being arrogant (and sometimes of being ignorant) and of not recognizing "the value of value pluralism." The fear of being the target of such accusations, together with the uncomfortable situation of charging colleagues whom one meets on a regular basis with doing substandard research, is probably an important reason why obscurantism continues to thrive.

Finally, obscurantism is sustained by the self-interest of non-obscurantist scholars. To be effective, an attack on obscurantism has to be well documented and well argued. Mere diatribes are pointless and sometimes counterproductive. Yet scholars have a greater personal interest in achieving positive results than in exposing the flaws of others, not only because of the reward system of science, but also because achieving positive results is intrinsically more satisfying. On grounds of self-interest, therefore, many scholars will hesitate to take time off from their main work and hope that someone else will do the cleaning up. The stage is set for another bad equilibrium. There are exceptions. Brian Barry, Robyn Dawes, and David Freedman performed disinterested public service by criticizing obscurantist work, line by line, equation by equation. The highly regarded economist Ariel Rubinstein has offered rare insider criticism of mainstream economics, commenting, for example, on "as-if rationality" that "it ultimately became clear to [him] that the phrase 'as if' is a way to avoid taking responsibility for the strong assumptions upon which economic models are founded." In his view, "economics is a culture and not a science."[22] Yet, important as they are, these are solitary efforts.

### Micro-mechanisms

I shall not say more about soft obscurantism, but attempt to sketch a *more modest and more robust* alternative to hard obscurantism. I shall add some

[22] Another critic, Robert Skidelsky, asserts that economics "is a form of post-Christian theology, with economists as priests of warring sects." While accurate, this statement will not worry the profession, since, unlike Rubinstein, Skidelsky is not a card-carrying mathematical economist. When Joseph Stiglitz, who *is* a (Nobel Prize-winning) mathematical economist, was asked at a private dinner party how economists can make repeated falsified claims without having their careers terminated, he reportedly answered: "I agree with you, but I don't understand why you are so puzzled. What you should be assuming is that – as is done by most economists – economics is really a religion. So why should you be puzzled by the fact that they cling to and never give up their views despite frequent falsification?" If Stiglitz really holds this view, he should be shouting it from the rooftops, not reserving it for dinner conversation.

detail to what I have said in earlier chapters, but the main purpose is to make a synthetic statement. After a brief comment on why one should not discard rational-choice models and statistical models altogether, I shall discuss micro- and macro-mechanisms, the importance of learning from the classical writers, and the importance of learning from history.

First, rational-choice theory and statistical analysis are indispensable tools, *when kept simple*. Although I cannot define simplicity, I can give a few pointers. Rational-choice theory is most likely to be useful when agents are not assumed to have well-defined beliefs about features of the world to which they have no direct access. These include macro-economic and macro-social facts, the preferences and beliefs of other agents, and the actions of people they do not know. Since most people do not know how much others donate to charity or whether others report their incomes correctly, explanations of their donations or reported incomes as a Nash equilibrium – the best response to the best responses of others – are implausible.

Regarding statistical modeling, my lack of personal competence forces me to cite the work of others. One warning against complexity states that "Where the medians and means (and basic cross-tabulations) don't persuade, the argument probably isn't worth making." According to another, a "statistical specification with more than three explanatory variables is meaningless." In this perspective, the numerous "large-N" cross-national regression analyses with up to a dozen independent variables are indeed meaningless. The most important thing to keep in mind, however, is the need to go beyond regressions and ask, in the words of David Freedman, "Does the model predict new phenomena? Does it predict the results of interventions? Are the predictions right?"

In Chapter 2, I argued for explanations by *mechanisms* rather than by laws. I shall first discuss micro-mechanisms and then macro-mechanisms.

Among the micro-mechanisms I have discussed, many have been studied, and some of them discovered, by behavioral economists. The following are especially important:

- loss aversion
- hyperbolic discounting
- decision myopia (choice bracketing)
- the sunk-cost fallacy
- altruistic punishment
- the hot–cold and cold–hot empathy gaps
- trade-off aversion
- anchoring in the elicitation of beliefs and preferences
- the representativeness and availability heuristics
- probability neglect
- duration neglect

- the certainty effect
- contrast and endowment effects
- motivated reasoning
- emotional reactions to unfair treatment
- flaws of expert judgments and of expert predictions
- magical thinking
- reason-based choice
- spurious pattern-finding.

Taken together, these and other mechanisms to be discussed shortly constitute an alternative to rational-choice theory. They do not form a coherent body, unified by deductive links. Instead, the accumulation of mechanisms – hardly a week goes by without a new effect being published on the internet – suggests that our beliefs, preferences, emotions, and choices are shaped by a bundle of unrelated mental quirks. If this is how we are, so be it. Yet, as I mentioned in Chapter 14, the fact that we *want* to be rational provides a counteracting force to the quirks. Moreover, when the stakes are high enough our self-interest may provide a corrective.[23] Learning over time can also weed out some anomalous behaviors, and selection weeds out those who are most prone to them. That being said, there is little doubt that many of the mechanisms are robust and hugely important in the explanation of social behavior.

The experimental setting has several artificial features that have to be kept in mind:

- The use of material (monetary) rewards and punishments is unusual outside the laboratory. In everyday life, we tend to express approval or disapproval, and seek out or avoid the target person.
- In some experiments designed to elicit how subjects make decisions under risk, they are *told* the possible outcomes and their probability. In everyday life, people have to figure out for themselves what the outcomes might be and how likely they are.
- Today, scholars are prohibited from conducting experiments with high-stake emotional charges (as in the Milgram experiments). Extrapolating from behavioral effects of the mild positive affect subjects feel when given candy or when discovering that the pay phone already has a coin in it may not be justified.
- Although the great care taken in many experiments to ensure subject–subject and experimenter–subject anonymity can be justified by the need to isolate intrinsic motives from socially induced ones, the infrequency of "anonymity in the wild" makes it hard to interpret the findings.

[23] Yet by using first-world research budgets to carry out high-stakes experiments in third-world societies, it has been found that people are willing to forego as much as a month's salary rather than being taken unfair advantage of.

- For practical reasons, an experimenter may prefer asking a subject how she *would react* if another subject behaved unfairly rather than observing how she *does react* to the same behavior.[24] The answer may not, however, be valid. Anger and indignation, for instance, are more easily triggered by actual harm done by others than by hypothetical harms. In experiments designed to compare the two situations, subjects impose lower levels of punishment in the hypothetical case.
- It is virtually impossible to recreate, inside the laboratory, the ongoing open-ended interactions that shape much social behavior. It is easy to *model* them as indefinitely iterated games, but the models suffer from the problems that I discussed earlier.
- Even when subjects are unobserved, the experimental situation may cause subjects to do what they think they are "supposed to do," either to behave in a ruthlessly competitive way or to be cooperative and altruistic. Simply labeling the situation as a "Wall Street Game" or a "Community Game" may influence behavior.

Behavioral economists are aware of these issues, and sometimes try to adduce evidence from "the wild" to show that their findings are not due to the artificial conditions of the experiments. It is important, though, to distinguish between *evidence* assessed by the hypothetico-deductive method (Chapter 1) and *anecdotes* that are merely consistent with the stipulated mechanisms. Let me give two examples.

To show the reality of the sunk-cost fallacy outside the laboratory, scholars sometimes cite the decisions to persist with the projects of the Anglo-French Concorde plane and the tunnel under the English Channel, or the pursuit to the bitter end of the wars that France and later the United States conducted in Vietnam. They rarely, however, take the further steps of (i) excluding alternative explanations and (ii) deducing and verifying additional implications. Concerning Concorde, for instance, Great Britain did want to scrap the project in 1964, when the anticipated cost of the plane had skyrocketed and the sonic boom proved so bad as to threaten its commercial viability. In the end, however, the British government decided to go ahead, because France might be awarded £200 million in compensation by the European Court if Britain cancelled unilaterally. In the

[24] For one, subjects might refrain from making unfair proposals, expecting that they would be rejected, thus making it difficult to verify whether and how often they are actually rejected. For another, eliciting responses to many actual proposals is more costly than to present the subject with a single list of hypothetical proposals. The first problem could be overcome by having subjects respond to computer-generated proposals, as long as they thought they were dealing with a real person. Given the anonymity of the experiments, this would be easy to achieve. A norm against this practice seems to be emerging in the behavioral economics community, however, because the experiments would cease to be reliable if the practice became known in the student populations from which most subjects are taken.

words of the biographer of Roy Jenkins, Minister of Aviation at the time, this possibility "made it more expensive to cancel than to carry on." It is, of course, possible that the *French* were victims of the sunk-cost fallacy.[25]

Let me also illustrate the issue at greater length, by citing a study on "norm theory" in which the authors assert a

> correlation between the perception of abnormality of an event and the intensity of the affective reaction to it, whether the affective reaction be one of regret, horror, or outrage. This correlation can have consequences that violate other rules of justice. An example that attracted international attention a few years ago was the bombing of a synagogue in Paris, in which some people who happened to be walking their dogs near the building were killed in the blast. Condemning the incident, a government official singled out the tragedy of the "innocent passers-by." The official's embarrassing comment, with its apparent (surely unintended) implication that the other victims were not innocent, merely reflects a general intuition: The death of a person who was not an intended target is more poignant than the death of a target.

The statement by the "government official" – it was in fact Raymond Barre, the Prime Minister at the time – is indeed consistent with the proposed explanation in terms of norm theory. It is also, however, consistent with an explanation in terms of an anti-Semitic prejudice. The evidence suggesting that Barre had an anti-Semitic bias includes his strong defense of Maurice Papon, notorious for his role in a round-up of French Jews in 1942, and a directive he signed in 1977 (later struck down by the Conseil d'État) that effectively cancelled anti-racist legislation from 1972. Moreover, Barre's actual comment was somewhat less innocuous than in the paraphrase of the authors. He referred to "the odious attack that intended to strike Jews on the way to the synagogue and that struck innocent French citizens crossing the street." In fact, the Jews in question were French too. In my view, this phrasing supports an explanation in terms of anti-Semitism. Although Barre may not have "intended" the implications that the Jewish victims were not innocent and that they were not French, prejudice often operates at an unconscious level (see Introduction to Part II).

Some behavioral economists do try to integrate laboratory findings and field studies in a more systematic way. In one example, which I propose as a model, a team of scholars conducted an experiment of the self-serving role of fairness with a follow-up study in the field. In the experiment, subjects were assigned to either the role of plaintiff or of defendant in a tort case and asked to negotiate a settlement. They were also asked to predict the award of the judge and to

[25] The American and French wars in Vietnam are also often cited as examples of the sunk-cost fallacy. An historian of the French war in Vietnam observes that civilian and military leaders claimed, as a "stock argument," that "Disengagement short of victory would insult the memory of the Frenchmen who had died defending the cause." Although similar statements were also made in the American context, my impression is that they were less frequent and/or less sincere.

assess what they considered a fair out-of-court settlement for the plaintiff, and were paid based on the accuracy of their answers. Plaintiffs predicted higher awards than defendants, and pairs of subjects were more likely to settle the more similar were their predictions and assessments. In other words, self-serving assessments of a high or low award made the subjects less willing to reach a negotiated agreement.[26] Moreover, the authors established that this was a causal effect and not a mere correlation, by running a variant of the experiment in which subjects made their assessments and predictions "behind a veil of ignorance," before they were assigned their roles. In that condition 6 percent of the pairs of subjects failed to settle, against 26 percent in the condition where the subjects knew what their interests were.

In the field study, the authors looked at negotiations between the teachers' union and the school boards in 500 school districts in Pennsylvania. Both sides insisted that wages be fair with respect to a reference group. The salary in the districts cited by the unions was on average about 2.4 percent, or $711 higher than the salary cited by the school board, suggesting a self-serving bias. Moreover, strikes occurred 49 percent more often in districts where the reference salary cited by the union was $1,000 higher than the reference salary cited by the school board, compared with districts where the reference salaries were the same. By itself, the study could not exclude that a third variable, such as the aggressiveness of the parties or the choice of extreme reference groups to justify industrial action, accounted for the difference. The laboratory experiments strongly suggested, however, that the choice of reference groups was self-serving and that it had a direct causal impact on the strikes.

Behavioral economics is not, of course, the only source for the study of micro-mechanisms. More traditional psychological approaches have discussed self-deception (Chapter 7), the effects of emotion on beliefs, preferences, and behavior (Chapter 8), as well as cognitive dissonance, reactance, and other mechanisms that I discussed in Chapter 9. I have also cited historians, novelists, and moralists, about whom more shortly, as sources of mechanisms. The caveat about the need to distinguish evidence from anecdotes obviously applies to these sources as well. In Chapter 1, I used the example of standing ovations on Broadway to illustrate how one may go beyond observing that this behavior is *consistent* with dissonance-reduction to argue that it is *explained* by that mechanism.

### Macro-mechanisms

By a macro-mechanism one might simply understand a micro-mechanism writ large, that is, triggered simultaneously in many people.[27] To explain, for

[26] In Chapter 12 I made a similar observation about child custody litigation.
[27] In fact, the standing ovations illustrate this case.

instance, the stability of highly hierarchical and unequal societies, we might cite the tendency of classes at the bottom of the hierarchy to be subject to adaptive preference formation. This aggregative view of macro-mechanisms is not very satisfactory, however. To see why, imagine a society in which brutal oppression causes the hatred of the subjects to dominate their fear. It does not follow, however, that they will take to arms. Not only is there little one individual can do, but also, crucially, he or she might not know how many others feel the same way and might join the fight. In what can be seen as an early statement of the idea of pluralistic ignorance, Seneca wrote that "A proposal was once made in the senate to distinguish slaves from free men by their dress; it then became apparent how great would be the impending danger if our slaves should begin to count our number."[28] There may be strength in numbers, but only if potential rebels have an idea about how many there are of them and assume that others, too, can estimate their numbers.

A more useful idea of a macro-mechanism is based on the *interaction and interdependence* of agents rather than on the aggregation of isolated individual reactions. Selection effects (Chapter 11), the "younger sibling" syndrome (Chapter 17), pluralistic ignorance and the "older sibling" syndrome (Chapter 22), as well as sequential unraveling and snowballing (Chapter 23), constitute macro-mechanisms in this sense. The following is an attempt to characterize them:

- Social agents have *preferences* and *beliefs*.
- Preferences can be defined either over *outcomes* (states of the world) or over *actions*.
- Preferences over actions can be induced either by preferences over outcomes (costs and benefits) or by non-consequentialist injunctions such as social norms, quasi-moral norms, and deontological moral norms.
- Preferences to do X can be *conditional*, and depend on the number of other agents whom the agent observes doing X (triggering either consequentialist or quasi-moral norms), or who are in a position to observe whether the agent does X (triggering social norms).

[28] Discussing a proposal to punish those who fail to show gratitude for a benefit, Seneca also wrote that "it is not advisable that it should be publicly known how many ungrateful men there are: for the number of sinners will do away with the disgrace of the sin." Similarly, it has been argued that publishing the number of unemployed removes the stigma of unemployment and lessens the incentive to find work. (Although people may be more affected by the proportion of unemployed among their friends and neighbors than by official statistics, the latter may also matter.) Thus, if an external shock throws a large number of individuals out of work, publishing the fact that there are so many of them might reduce their incentive to get back to work, so that unemployment might persist even when economic conditions improve (hysteresis).

- The preferences and beliefs of an individual jointly induce actions. At the same time, as noted, the preferences may themselves depend on observation-induced beliefs.
- Whereas an agent cannot observe the beliefs and preferences of other agents, he may be able to observe their actions (including their statements), or inactions, and try to infer their beliefs and preferences from them.
- These inferences, which may well be wrong, can then serve him as premises for further actions. When many people do so, their actions may or may not confirm the premises.
- Even when the agent cannot observe individual actions, he may be able to observe the aggregate outcome of actions or the average propensity to act, and use this information as the basis for his further action.
- At the same time as the agent is forming beliefs about other agents, he knows that they are forming beliefs about him, on the basis of their observations of what he does.
- Because he may have preferences regarding these beliefs (he may not want to be thought badly of), his beliefs about the beliefs of others about him can shape his behavior.
- Authorities can change individual preferences directly by the use of punishments and rewards. They can also do so indirectly, by providing aggregate or individualized information that trigger conditional preferences.

Taken together, these relations create *networks of nested beliefs and preferences* that may explain actions or decisions to abstain from action.

### *The classics*

Throughout this work I have constantly cited, often at great length, writings by classical authors, men and two women (Jane Austen and George Eliot) who wrote about social affairs before the creation of specialized academic disciplines of social science. A priori, it makes sense to draw on their insights. There is no reason why the last century or the last decade should have a privileged status in the generation of psychological and sociological mechanisms. If we ignore the classics, we do so at our loss and our peril.

True, if social science, like the natural sciences, were based on *laws*, there would be little reason to read the classics, except from the perspective of the history of ideas. Alfred Whitehead said, "A science that hesitates to forget its founders is lost." His statement is too strong – Darwin is still worth reading – but essentially correct. Once the findings of the past have been rendered into easily assimilated textbook material, there is no reason to revisit the often stumbling and confused first efforts, except, to repeat, if those efforts are what we want to understand. The social sciences, by contrast, *progress by the*

*accumulation of mechanisms*. When a new mechanism is added to the toolbox, it does not replace previous ones.

I shall say a few words about some of the classics I have cited.

*Seneca* (the Younger), the richest man in the Roman world of his time, was the tutor of Nero and at the age of sixty-nine killed himself at Nero's order. *Michel de Montaigne* was mayor of Bordeaux during the French wars of religion and had close relations with Henri de Navarre (the later Henri IV), but mostly lived the life of a landed gentleman.[29] *Blaise Pascal* had one of the greatest intellects of all time, with interests ranging from mathematics and physics to Jansenist theology. The *Duc de la Rochefoucauld* was a military man, deeply involved in the "Fronde," an intrigue of French nobles around 1650. *Jean de la Bruyère* was a tutor of princes and princesses at the court of Louis XIV. *Samuel Johnson* was the most distinguished man of letters of his time, and the object of perhaps the most famous biography ever written. *David Hume* never had a fixed profession, but acquired considerable wealth from the sales of his *History of England*. *Edward Gibbon* was member of parliament, with independent means that allowed him to focus on his work. *Adam Smith* was first a professor of moral philosophy in Glasgow, then a tutor of a young nobleman, and finally commissioner of customs in Scotland. Hume, Gibbon, and Smith traveled extensively on the continent, notably in France, where they met the leading intellectuals of the day. *Jeremy Bentham*, too, had close relations with a circle of French politicians at the time of the Revolution; in England, he was deeply and constantly involved in various reform projects. *Jane Austen* lived her short life embedded in the village lives she describes, capturing the finest nuances of behavior with the attention of an entomologist. *Stendhal* (Henri Beyle, by his real name) was active in Napoleon's Italian, German, and Russian campaigns and later served as a diplomat. *Alexis de Tocqueville* was a lawyer by training, a close observer of the French revolutions of 1830 and 1848, later a member of the French National Assembly and briefly Minister of Foreign Affairs. *Marcel Proust*, author of the greatest novel of the twentieth century, spent much of his life frequenting salons, not only observing the goings-on in microscopic detail, but also identifying underlying psychological mechanisms.

I offer these less-than-thumbnail descriptions to make the point that by virtue of their wide-ranging experience, these writers had a deep understanding of human nature and of social life. Some traveled widely, were active in political and military affairs, and knew danger. Others lived cloistered lives, but used their powers of observation and analysis to identify mechanisms that

[29] He was deeply familiar with Latin and the Roman classics, notably Seneca and Plutarch. His father ensured that as a child he would be addressed only in Latin, even by members of his family and the servants, chosen because they spoke that language.

transcend the villages or salons. Some of their insights were rediscovered by social scientists centuries or millennia later. Examples include pluralistic ignorance (Seneca, Tocqueville), the "white bear effect" (Montaigne), the endowment-contrast effects (Montaigne, Hume), the misinterpretation of one's own feelings (Austen), adaptive preferences (the French moralists, Tocqueville), focal points (Pascal), magical thinking (Proust), the free-rider problem in information-gathering (Bentham), other forms of the Prisoner's Dilemma (Hume, Marx, Tocqueville), cognitive dissonance (Montaigne, Proust). There is an obvious risk of overinterpreting such precedents. When a writer makes a remark in passing without appreciating its implications and importance, one should be careful in attributing priority.[30] Attempts to find anticipations of hyperbolic time discounting in Hume and Adam Smith are strained. Yet some of the ideas I cited are stated very precisely, and the brevity of the statements is explained by the fact that their authors were not concerned with making a contribution to social science.

Other insights seem to have escaped the attention of academic scholars. What I have called the psychology of tyranny, discussed extensively by Seneca, Gibbon, and Tocqueville, has not been a topic for modern scholars, perhaps because they tend to assume that preferences are stable. Nor have psychologists expanded on Proust's observation on the transmutation of motives: "our imagination . . . substitutes for our primary motives alternative motives that are more acceptable." The idea that the spontaneous action tendency of revenge might be "two eyes for an eye" (Seneca, Adam Smith) rather than one eye does not seem to have caught the attention of the behavioral economists who study punishment in the laboratory. Nor have they taken up a central idea in Seneca and the French moralists, "those whom they injure, they also hate." The distinction between wanting to *make* something the case and wishing something to *be* the case (Seneca, Adam Smith) has been lost, as has Tocqueville's distinction between cognitive and motivational myopia. Advocates of bicameralism could usefully have pondered Gibbon's observation that passions can undermine the precautions people take against them. Now, as it is difficult to prove a negative, especially since my knowledge of the literature is limited, some or all of the claims in this paragraph may be wrong. I think I am on safe ground, however, in asserting that these ideas do not have the place in modern scholarship they deserve.

[30] In 1831, a Scottish landowner, Patrick Matthews, published a book *On Naval Timber and Arboriculture* in which, Darwin wrote, "he briefly but completely anticipates the theory of Nat. Selection." The fact that this proto-theory of natural selection was relegated to an appendix in the book shows that the author did not understand the importance of his discovery.

### The historians

I believe the best training for any social scientist is to read widely and deeply in history, choosing works for the intrinsic quality of the argument rather than the importance or relevance of the subject matter. Here are some suggestions:[31] James Fitzgerald Stephen, *A History of the Criminal law of England*; E. P. Thompson, *The Making of the English Working Class*; G. E. M. de Ste Croix, *The Class Struggles in the Ancient Greek World*; Joseph Levenson, *Confucian China and its Modern Fate*; Paul Veyne, *Le pain et le cirque* and a follow-up collection of essays, *L'empire gréco-romain*; G. Lefebvre, *La grande peur*; Keith Thomas, *Religion and the Decline of Magic*; Tocqueville's *L'ancien régime et la Révolution*; two books on the *ancien régime* by Marcel Marion, volume I of his *Histoire financière de la France depuis 1715* and *Machault d'Arnouville*; Gordon Wood, *The Radicalism of the American Revolution*; Jean Egret, *La pré-révolution française*; Alan Taylor, *The Internal Enemy*; two books on very different topics by Marc Bloch, *Les rois thaumaturges* and *Les caractères originaux de l'histoire rurale française*; two outstanding books on the Vietnam War, H. R. McMaster, *Dereliction of Duty* and L. Gardner, *Pay any Price*; Paul Langford, *Public Life and the Propertied Englishman 1689–1798*; Martin Ostwald, *From Popular Sovereignty to the Sovereignty of Law*; J. R. Pole, *Political Representation in England and the Origins of the American Republic*; J. Uglow, *In These Times: Living in Britain through Napoleon's Wars, 1793–1815*; Geoffrey Parker, *Imprudent King*; and two caustic books by Peter Novick, *That Noble Dream* (on the search for objectivity by American historians) and *The Holocaust in American Life*. What these writers and others of their stature have in common is that they combine utter authority in factual matters with an eye both for potential generalizations and for potential counterexamples to generalizations. By virtue of their knowledge they can pick out the "telling detail" as well as the "robust anomaly," thus providing both stimulus and reality check for would-be generalists.

The same is true for authors of good "case studies," among which one of the greatest remains Tocqueville's *Democracy in America*. Although it does not fit neatly into the category, I would also include Joseph Schumpeter, *Capitalism, Socialism, and Democracy*. A seemingly eccentric but, I believe, compelling candidate is Arthur Young's *Travels in France*, covering the years 1787, 1788, and 1789. These are "character portraits" of whole societies or regimes, all of them with a comparative perspective. Marc Bloch, *La société féodale*, also belongs here. Alexander Zinoviev's *The Yawning Heights* is not exactly a character portrait of postwar Soviet Communism, but a caricature in the good

[31] They are somewhat parochial, citing works only in the two languages I master well.

sense of the word – eliminating inessentials and isolating core features by exaggerating them. It is usefully supplemented by F. Stufford, *Red Plenty,* and by a study of prewar Communism by S. Fitzpatrick, *Everyday Stalinism*. The trilogy by Richard Evans on the Third Reich did for the specific regime of Nazism what Robert Paxton did in *What Is Fascism?* for the more generic regime. Richard Bosworth's *Mussolini* and *Mussolini's Italy*, if read in conjunction with Evans's books, provide striking insights into the difference between a regime whose evil, while real, was largely low-grade and one that was evil to the core.

Two multi-volume books are in a class by themselves, Hume's *History of England* and Gibbon's *Decline and Fall of the Roman Empire* (already included among "the classics"). Hume of course was not mainly an historian, and took most of his factual material from secondary sources. Yet the appendices to the six volumes show the care he took to hold various accounts up against each other and to examine their intrinsic plausibility, using some of the same methods he deployed in his essay on miracles. The work is mainly important, however, as a pioneering effort in political psychology, equaled, maybe surpassed, in his time only by Gibbon, who *was* a professional historian. Both Hume and Gibbon were open to the variety and complexity of human motivations, as I hope will be clear from the passages I cite from them. They were also admirably, almost programmatically, free of cant. Gibbon's irony, like that of Peter Novick, is especially refreshing.

### Putting it all together

Good scholars need intelligence, creativity, persistence (*Sitzfleisch*), and intellectual honesty. (Luck, too, is useful.) Outside mathematics and physics, a high level of intelligence is not essential, although a modicum is obviously necessary. Creativity seems to depend both on the innate capacity of the unconscious to form associations, which cause the solution to a problem to appear when you wake up in the morning, and on the accumulation of elements between which those associations might be made. That accumulation in turns depends on a wide and broad reading of the classics and of history. The classics can provide explicit mechanisms, often in lapidary form. Historians often provide implicit or potential mechanisms, in addition to showing us the varieties of human behavior and social organization. Psychology and behavioral economics can refine the mechanisms and transform them into testable hypotheses, as well as coming up with ideas that nobody has thought of. Persistence is needed for the necessary attention to detail. It is too much to ask that scholars should have "the infinite capacity for taking pains" that has been used as a definition of genius, but they should use shoe leather. Intellectual honesty may not matter much in mathematics and physics, since formal

proofs and replicable experiments do not depend on the possession of that quality. Honesty (and modesty) is vital, however, in disciplines where the constraints created by deductive logic and hard facts are lacking. If someone asked me how to acquire it, I would say: read Montaigne.

### Bibliographical note

I first engaged in sustained criticism of hard obscurantism in a review essay of R. Bates *et al*., *Analytic Narratives* (Princeton University Press 1998), published as "Rational-choice history: a case of excessive ambition?" *American Political Science Review* 94 (2000), 685–95, followed by a reply from the authors. (Later, I dropped the question mark.) More recently I discussed hard obscurantism in "Excessive ambitions," *Capitalism and Society* 4(2) (2009), Article 1, and both the hard and soft varieties in "Hard and soft obscurantism in the humanities and social sciences," *Diogenes* 58 (2102), 159–70. The first of these was followed by hard-hitting replies by eminent practitioners of, respectively, rational-choice modeling and data analysis, Pierre-André Chiappori and David Hendry. The founding article of bullshittology is H. Frankfurt, "On bullshit," *Raritan Quarterly Review* 6 (1986), 81–100. A useful analysis of soft obscurantism is F. Buekens and M. Boudry, "The dark side of the loon: explaining the temptations of obscurantism," *Theoria* 81 (2014), 126–42. The analysis of Baudelaire's poem is in R. Jakobson and C. Lévi-Strauss, "*Les Chat*s de Charles Baudelaire," *L'Homme* 2 (1962), 5–21. The list of defense mechanisms is taken from G. Vaillant, *Ego Mechanisms of Defense* (Washington, DC: American Psychiatric Association Press 1992). The remark on false windows in the analogy between empire and monotheism is from P. Veyne, *L'empire gréco-romain* (Paris: Seuil, 2005), p. 336. The study of marriage and migration patterns in South India is M. Rozensweig and O. Stark, *Journal of Political Economy* 97 (1989), 905–26. Kenneth Arrow's remark on social norms is in his "Political and economic evaluation of social effects and externalities," in M. Intriligator (ed.), *Frontiers of Quantitative Economics* (Amsterdam: North-Holland, 1971), pp. 3–25. The study of endogenous time discounting is G. Becker and C. Mulligan, "The endogenous determination of time preference," *Quarterly Journal of Economics* 112 (1997), 729–58. The study of endogenous altruism is C. Mulligan, *Parental Priorities and Economic Inequality* (University of Chicago Press, 1997). A study of endogenous risk attitudes is I. Palacios-Huerta and T. Santos, "A theory of markets, institutions, and endogenous preferences," *Journal of Public Economics* 88 (2004), 601–27. A study of revolutionary transitions is D. Acemoglu and J. Robinson, "A theory of political transitions," *American Economic Review* 91 (2001), 938–63. The article by Gordon Tullock they refer to is "The paradox of revolution," *Public Choice* 11 (1971), 89–99. The mixed-strategy

analysis of the "Kitty Genovese" case is in A. Dixit and S. Skeath, *Games of Strategy* (New York: Norton, 2004). Two outstanding books by David Freedman are *Statistical Models* (Cambridge University Press, 2005) and *Statistical Models and Causal Inference: A Dialogue with the Social Sciences* (Cambridge University Press, 2010). The former reproduces in their entirety and criticizes four articles from leading social-science journals. The latter includes his article on "Statistical models and shoe leather." Another important contribution along the same lines is C. Achen, "Towards a New Political Methodology," *Annual Review of Political Science* 5 (2002), 423–50. The assessment of the time needed to learn optimal rules by trial and error is T. Allen and C. Carroll, "Individual learning about consumption," *Macroeconomic Dynamics* 5 (2001), 255–71. The cited study of data mining and out-of-sample testing is A. Inoue and L. Kilian, "In-sample or out-of-sample tests of predictability: which one should we use?" *Econometric Reviews* 23 (2004), 371–402. An introduction to controlled randomization is A. Banerjee and E. Duflo, *Poor Economics* (New York: PublicAffairs, 2012). An introduction to instrumental variables is A. Sovey and D. Green, "Instrumental variables estimation in political science: a reader's guide," *American Journal of Political Science* 55 (2010), 188–200. A devastating account of the life and work of Bruno Bettelheim is R. Pollak, *The Creation of Dr. B.: A Biography of Bruno Bettelheim* (New York: Touchstone Books, 1997). A critical and historical discussion of attachment theory is M. Vicedo, *The Nature and Nurture of Love: From Imprinting to Attachment in Cold War America* (University of Chicago Press, 2013). The passage by John Bowlby is cited after this book. For Freud's misogyny, see his *Gesammelte Werke* (Frankfurt am Main: Fischer, 1947), vol. XII, p. 176, vol. XV, pp. 142, 144. An excellent study of the intellectual and therapeutic shortcomings of psychoanalysis is J. van Rillaer, *Les illusions de la psychanalyse* (Brussels: Éditions Mardaga, 1980). The study of how psychoanalysis hindered the treatment of drug addicts is J.-J. Deglon, "Comment les théories psychanalytiques ont bloqué le traitement efficace des toxicomanes et contribué à la mort de milliers d'individus," in C. Meyer (ed.), *Le livre noir de la psychanalyse* (Paris: Éditions des Arènes, 2010), pp. 516–41. The negative impact of the theory on the treatment of schizophrenia and autism is discussed in V. Gueritault, "Les mères, forcément coupables," *ibid*., pp. 544–72. The cited passage on the repressed memory syndrome is from E. Loftus, "Our changeable memories: legal and practical implications," *Nature Reviews Neuroscience* 4 (2003), 231–4. My comments on the Vietnam War draw on H. R. MacMaster, *Dereliction of Duty* (New York: Harper, 1997), L. Gardner, *Pay any Price: Lyndon Johnson and the Wars for Vietnam* (Chicago: Elephant Paperbacks, 1997), and Kai Bird, *The Color of Truth: McGeorge Bundy and William Bundy* (New York: Touchstone Books, 1998). Critics of flawed arguments about the effect of handguns and

the death penalty include I. Ayres and J. Donahue, "Shooting down the 'more guns, less crime' hypothesis," *Stanford Law Review* 55 (2003), 1193–1312, and J. Donohue and J. Wolfers, "Uses and abuses of empirical evidence in the death penalty debate," *Stanford Law Review* 58 (2005): 791–846. The comment on John Lott's argument for handguns is by Hashem Dezhbakhs, as cited by Ayres and Donahue, who add that it "is equally applicable to the debate over capital punishment." On the dangers and costs of mechanical diversification of assets, see A. Bhidé, "In praise of more primitive finance," *The Economist's Voice*, February 2009, pp. 1–8. The argument about the roles of the left and right hemispheres are from V. S. Ramachandran and S. Blakeslee, *Phantoms in the Brain* (New York: Quill, 1998). On the question why string theory has acquired great prestige without making confirmed predictions, see L. Smolin, *The Trouble with Physics* (Boston: Houghton Mifflin, 2007). On foot binding, see G. Mackie, "Ending footbinding and infibulation: A convention account," *American Sociological Review* 61 (1996), 999–1017. Invaluable public service in debunking soft obscurantism was performed by Robyn Dawes, *House of Cards: Psychology and Psychotherapy Built on Myths* (New York: The Free Press, 1996), and by Brian Barry, *Culture and Equality: An Egalitarian Critique of Multiculturalism* (Cambridge MA: Harvard University Press, 2002). Ariel Rubinstein's insider criticism of economic theory is in *Economic Fables* (Cambridge: Open Book, 2012). His observation that economics is a culture rather than a science is in his "Comment on neuroeconomics," *Economics and Philosophy* 24 (2008), 485–94. The remark by Joseph Stiglitz is reported in A. Bilgrami, "Truth, balance, and freedom," in A. Bilgrami and J. Cole (eds.), *Who's Afraid of Academic Freedom?* (New York: Columbia University Press, 2015), pp. 20–1. The comment on the Concorde project is in J. Campbell, *Roy Jenkins* (London: Jonathan Cape, 2014), p. 248. The comment on the behavior of the "French official" is from D. Kahneman and D. Miller, "Norm theory," *Psychological Review* 93 (1986), 136–53. The articles on fairness and bargaining that I held up as models are summarized in L. Babcock and G. Loewenstein, "Explaining bargaining impasse: the role of self-serving biases," *Journal of Economic Perspectives* 11 (1997), 109–26.

# Index