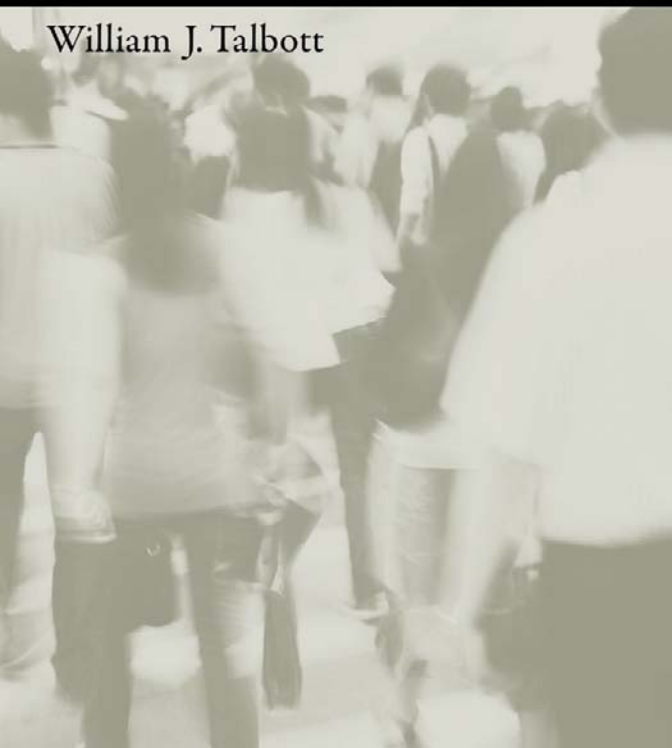
Human Rights and Human Well-Being

William J. Talbott



HUMAN RIGHTS AND HUMAN WELL-BEING

OXFORD POLITICAL PHILOSOPHY

GENERAL EDITOR: SAMUEL FREEMAN,
UNIVERSITY OF PENNSYLVANIA

Oxford Political Philosophy publishes books on theoretical and applied political philosophy within the Anglo-American tradition. The series welcomes submissions on social, political, and global justice, individual rights, democracy, liberalism, socialism, and constitutionalism.

N. Scott Arnold
Imposing Values: An Essay on Liberalism and Regulation

Peter de Marneffe
Liberalism and Prostitution

William J. Talbott
Human Rights and Human Well-Being

HUMAN RIGHTS AND HUMAN WELL-BEING

William J. Talbott

OXFORD
UNIVERSITY PRESS
2010

OXFORD
UNIVERSITY PRESS

Oxford University Press, Inc., publishes works that further
Oxford University's objective of excellence
in research, scholarship, and education.

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Published by Oxford University Press, Inc.
198 Madison Avenue, New York, New York 10016

www.oup.com

Library of Congress Cataloging-in-Publication Data
Talbott, W. J.
Human rights and human well-being / William J. Talbott.
p. cm.—(Oxford political philosophy)
Includes bibliographical references.
ISBN 978-0-19-517348-2
1. Human rights—Philosophy. 2. Human rights—Moral and ethical aspects.
I. Title.
JC571.T14445 2005
323.01—dc22 2009045410

9 8 7 6 5 4 3 2 1

Printed in the United States of America
on acid-free paper

In Memory of Robert Nozick,
Teacher and Friend

This page intentionally left blank

I have the innate duty . . . so to affect posterity through each member of the sequence of generations in which I live, simply as a human being, that future generations will become continually better . . . and that this duty may thus rightfully be passed on from one generation to the next. . . . Without this hope for better times the human heart would never have been warmed by a serious desire to do something useful for the common good.

—Immanuel Kant

This page intentionally left blank

Acknowledgments

This is the second of two volumes on human rights that attempt to answer the question "Which rights should be universal?" and to explain why they should be. I have already thanked those who helped with the first volume. I have many more people to thank for helping me to complete this project.

Let me begin by thanking my students. I was fortunate to be able to teach an early draft of this book in a seminar at the University of Washington in spring 2005. I am grateful to the students in that seminar for much valuable feedback: Ben Almassi, Daniel Baker, Ron Belgau, Erika Dahlstrom, Ethel Evans-Elison, Jeremy Fischer, Leann Haggard, Philip McGrane, Scott McMahon, Brandon Morgan-Olsen, Kelly O'Connell, Dustin Pearson, Gabriela Remow, Benjamin Robbins, Christi Siver, and Edward Wolcher. I am also grateful to students in my course on liberty for feedback on my consequentialist account of rights against paternalism, especially to Adam Caldwell, John Gresham, Quinn Rotchford, and Shawn Stevenson.

One advantage of writing two volumes on human rights is that I have been able to take account of responses to the first volume in the second. I have been fortunate to have received many cogent criticisms of the first volume. At an APA Pacific Division session in March 2006, my critics included Carol Gould, James Nickel, and David Reidy. Gould and Reidy's criticisms and an additional critical review by Jeppe von Platz were published with my replies (Talbott 2008). At a symposium at the University of Washington in April 2006, I was fortunate to have Henry Shue and Kok-Chor Tan as my critics. Henry Shue reprised his role at an APSA session in August 2006, where he led a formidable lineup of critics that included Brooke Ackerly, Charles Beitz, and Jack Donnelly. Their criticisms and my replies were subsequently published (Talbott 2007). Finally, I received criticisms from Christopher Knapp at an APA Eastern Division session in December 2007. At all of these sessions, I also received valuable feedback from members of the audience.

In addition to opportunities to respond to criticisms of the first volume, I was fortunate to receive a number of invitations that allowed me to try out some parts of this second volume before publishing it. I presented an earlier version of the first three chapters at the NYU Colloquium in Legal, Social, and Political Philosophy in September 2006. I am especially grateful to Jeremy Waldron for his sympathetic presentation of my view, to Ronald Dworkin and

Samuel Freeman for their criticisms, and to Thomas Nagel for helpful discussion. An earlier version of the first chapter was the basis for a presentation at the Conference on Human Rights and the New Global Order at the Kennedy School of Government in May 2008 and at an APA Pacific Division symposium on Consequentialist Foundations for Liberal Rights in April 2009, where I especially benefited from the comments of Samuel Freeman and Richard Arneson.

Earlier versions of chapters 12 and 13 on rights against paternalism were presented in four venues: first, at a colloquium at the University of Alberta in September 2003, where I especially benefited from the comments of Martin Tweedale and Karen Houle; then at a GALA seminar at the Kadish Center for Morality, Law & Public Affairs at UC Berkeley in March 2004, where I especially benefited from the comments of Samuel Scheffler, Jay Wallace, Meir Dan-Cohen, and Jodi Halpern; at an interdisciplinary colloquium on privacy rights at the University of Utah in April 2004; and at the UW Center for Statistics and Social Sciences in May 2004.

In addition, many people have given me comments on one or more chapters, including Michael Blake, Jeff Clausen, Jacques Corriveau, Samuel Freeman, Stephen Gardiner, Sara Goering, Lauren Hartzell, Eunjung Katherine Kim, Brad McHose, Adam Moore, Liam Murphy, Angela Smith, and Richard Zerbe. My discussion of disability rights is indebted to comments from Holly Siegrist and Scarlett Mai. I am grateful to all of these readers, and especially to Jamie Mayerfeld, who gave me substantive comments on the entire manuscript. I am pleased that Elizabeth Ashford, who read the complete manuscript as a reviewer for Oxford University Press and gave me useful feedback, allowed Oxford University Press to remove the cloak of anonymity so that I can thank her here.

I was able to do a great deal of the writing in 3 months at the UW's Helen Riaboff Whiteley Center over the period 2007 to 2010. I greatly appreciated the opportunity to work in an environment so conducive to thought.

This book would not have existed—at least, not in anything like its current form—without the support at crucial stages of Liam Murphy and Thomas Nagel. I am also grateful to Samuel Freeman for including it in his series of works in political philosophy. I thank my editor at Oxford University Press, Peter Ohlin, for his support and his patience over many years and the other members of the OUP production and marketing team, Liz Smith, Linda Donnelly, Stephanie Attia, Elyse Turr, and freelance copyeditor Mary Anne Shahidi. Thanks also to my daughter Rebecca for help with legal research and to my daughter Kate for proofreading. My deepest debt of gratitude is owed to my wife, Judy, whose love and understanding have sustained me through the entire process.

This book really began in an undergraduate political philosophy class taught by T. M. Scanlon in the spring of 1968. It was in that course that I read J. S. Mill's *On Liberty* and a draft of Scanlon's "A Theory of Freedom of Expression" for the first time. Although Scanlon was not a consequentialist,

both he and Mill provided models of indirect arguments for autonomy rights—that is, arguments for autonomy rights not grounded in the intrinsic value of autonomy. The following year I did an independent study with Scanlon on the manuscript of John Rawls's *Theory of Justice*. Again I was struck by Rawls's attempt to ground the autonomy rights of his special conception of justice indirectly in what seemed to me to be the well-being considerations of his general conception. This work with Scanlon was formative for my philosophical outlook, and I am deeply appreciative of his contribution to my thinking.

I first began to articulate my own attempt to ground individual rights indirectly on considerations of well-being when, as a graduate student, I took Robert Nozick's course in political philosophy in the fall of 1972. In that course, Nozick taught the manuscript of his philosophical defense of libertarianism, *Anarchy, State, and Utopia*. I wrote a term paper in which I criticized Nozick's libertarian position from what, in retrospect, I can see was a kind of Millian-Rawlsian indirect consequentialist viewpoint. The paper was a distant precursor of the indirect consequentialist position that I defend in this book. It led to lots of good philosophical give-and-take. Anyone who knew Nozick can tell you that I could not have had a more trenchant critic. He was also a warm human being and a source of unstinting encouragement.

I always knew that someday I would revise that term paper. Finally, 37 years later, I am ready to turn in the final draft. Sadly, my professor is no longer able to accept late work.

This page intentionally left blank

Contents

This page intentionally left blank

HUMAN RIGHTS AND HUMAN WELL-BEING

This page intentionally left blank

ONE

The Consequentialist Project for Human Rights

In this, the second of two volumes, I continue the project of explaining which rights should be universally guaranteed to all normal human adults by governments everywhere.[1] In the first volume I focused on what I regard as the basic human rights. In this book I discuss both basic and nonbasic human rights and explain more fully why the rights I discuss, both basic and nonbasic, should be universal. I have written this book to stand on its own, so that it is not necessary to have read the first volume before reading this one.

My goal is to contribute to an important explanatory project in political philosophy. In this chapter I say what the project is and provide an overview of how I propose to contribute to it. The first volume dealt extensively with the metaphysics and the epistemology of moral belief. In this chapter, I review that discussion briefly and then, in chapters 7 and 8, I develop the epistemology more fully.

Mill's and Rawls's Consequentialist Projects

Perhaps the best way to introduce the project of this book is to do so historically. The project began in the 1850s with J. S. Mill's *On Liberty* [1859].[2] Mill's book was to be a new kind of defense of a package of autonomy rights, including rights to freedom of thought and discussion, freedom of the press, freedom of association, and freedom from paternalism. Mill was not the first philosopher to defend a package of autonomy rights. Locke and Kant, among many others, had defended such rights long before Mill. What made Mill's defense of them distinctive was that he did not begin by assuming such rights or by assuming that they were to be justified by the intrinsic value of autonomy. He intended to show that a package of autonomy rights could be justified on utilitarian grounds—that is, on the basis of the contribution to overall well-being that would result from the government's legally enforcing them.

Would the rights be absolute, so that no exceptions could ever be justified? Though Mill sometimes wrote in a way that suggested the rights should be absolute, from his discussion of examples it is clear that he did allow for exceptions.[3] This is not surprising, because even most nonconsequentialists allow that rights have some exceptions.[4]

At the time that Mill was writing, it was generally assumed that the only kind of rights that a utilitarian could justify were rights the government should infringe whenever it thought that infringing them would maximize overall utility. Call this sort of right an *act utilitarian* right. Mill was aiming to defend rights much stronger than this. Even if autonomy rights should not be absolute, Mill would argue on utilitarian grounds that autonomy rights ought to be stronger than act utilitarian rights—strong enough, that is, that a government could not justify infringing them simply because the government thought that the infringement would maximize overall utility. Indeed, because Mill regarded autonomy rights as protections against not only government tyranny, but also tyranny by a majority, he clearly intended that they be strong enough not to be overridden by a simple majority. I refer to rights of this kind as *robust rights*. Because robust rights need not be absolute, there is no presumption that they can never be overridden, only that what is necessary to override them is significantly more demanding than what is necessary to override act utilitarian rights. Thus, for example, there is no paradox in thinking that in enforcing such rights the courts would sometimes have to prohibit actions that the government believes will maximize overall utility or to invalidate legislation adopted by a majority vote in the legislature.

There was one final element in Mill's account. Mill believed that there were utilitarian grounds for holding that at least some autonomy rights should be *inalienable*—that is, that at least some autonomy rights generate limits on the rights bearer's autonomy to trade or surrender those very rights. For example, Mill argued that people should not be free to enter slavery contracts ([1859], 115).

Because Mill was a utilitarian, he qualifies as a *consequentialist* in the sense in which I use the term: A *consequentialist* about a given normative domain is someone who believes there is an explanation of that domain in terms of some measure (perhaps a distribution-sensitive measure) of nonmoral good. For Mill, the measure of nonmoral good was utility maximization. If the relevant measure is a (perhaps distribution-sensitive) measure of *well-being*, the view is *welfare consequentialist*, or *welfarist*. As a utilitarian, Mill was a welfare consequentialist about all of morality.

Utilitarianism is a maximizing view, which makes it a *teleological* view. My version of welfare consequentialism about human rights is not a maximizing view, because, on my account, both the amount and the distribution of well-being matter.[5]

Mill was the first person to attempt to give a consequentialist explanation of why governments should guarantee to all normal adults a package of robust, inalienable autonomy rights. I refer to this project as the *consequentialist project for autonomy rights*.

It is generally agreed today that Mill's attempt to carry out the consequentialist project for autonomy rights failed. One of the philosophers most responsible for this verdict is John Rawls. A little over 100 years after the publication of *On Liberty*, John Rawls wrote *A Theory of Justice*. Let me refer

to the author of that book as the early *metaphysical* Rawls, to distinguish him from the later *political* Rawls who would disavow parts of it.[6]

Because the early metaphysical Rawls was writing in the shadow of J. S. Mill, he began his book in a way that was best calculated to separate himself from Mill:

> Justice is the first virtue of social institutions, as truth is of systems of thought. A theory however elegant and economical must be rejected or revised if it is untrue; likewise laws and institutions no matter how efficient and well-arranged must be reformed or abolished if they are unjust. Each person possesses an inviolability founded on justice that even the welfare of society as a whole cannot override. For this reason justice denies that the loss of freedom for some is made right by a greater good shared by others. It does not allow that the sacrifices imposed on a few are outweighed by the larger sum of advantages enjoyed by many. (1971, 3–4)[7]

The rhetorical force of his introduction tended to obscure how much the early metaphysical Rawls had in common with Mill—especially that metaphysical Rawls was attempting to bring to a successful completion the consequentialist project for autonomy rights.[8]

Unlike Mill, metaphysical Rawls did not try to give a consequentialist account of all of morality. His more modest goal was a consequentialist account of the justice or injustice of the basic institutions of society, especially the rights and duties established by law (including the constitution) and enforced by the coercive power of the state. Because our legal rights and duties provide a framework that defines our entitlement to the distribution of the benefits and burdens of social cooperation, metaphysical Rawls referred to his theory as a theory of *distributive justice*. Metaphysical Rawls believed that it was possible to develop the theory of distributive justice without working out a theory of *corrective justice* (i.e., a theory of the justice or injustice of punishment and other legal sanctions) if he focused on an ideal conception of justice for a society on the assumption of strict compliance with the just laws (1971, 8). On the assumption of strict compliance, it was not necessary for the early Rawls even to address issues of corrective justice.

The early Rawls's theory of distributive justice is almost universally identified with his two principles of justice, plus the priority rule that gives the first principle that specifies the basic autonomy rights (the Liberty Principle) lexical priority over the second (the Difference Principle) (1971, 302–303).[9] The Liberty Principle requires government protection of a package of robust, inalienable autonomy rights, very similar to Mill's package of rights, plus democratic rights. I refer to the combination of autonomy rights and democratic rights as *liberal rights*.[10]

Because Rawls's special theory of justice accords lexical priority to the rights covered by the Liberty Principle over considerations of well-being

(which are included in the Difference Principle), it is almost universally regarded as a nonconsequentialist theory of distributive justice, and the early Rawls himself seems to have regarded it as such (1971, 11). It is undeniable that the later *political* Rawls's (1993) account of the two principles (and the priority of the first over the second) is nonconsequentialist, but it is often overlooked that the early metaphysical Rawls regarded those two principles as a special case of a general conception of distributive justice that contained only a single, consequentialist principle, which I refer to as Rawls's *maximin expectation principle*—roughly to maximize the expectations of the least advantaged group.[11] Thus, for all his attempts to distance himself from Mill, if the early metaphysical Rawls had been successful, he would have succeeded in carrying out the consequentialist project for liberal rights, including both autonomy rights and democratic rights—that is, the project of providing a consequentialist explanation (in terms of his maximin expectation principle) of why governments should guarantee a package of robust, inalienable autonomy and democratic rights to all normal adults (i.e., a derivation of the Liberty Principle and of its lexical priority over the Difference Principle).[12]

Unlike Mill, the early Rawls never envisioned providing a consequentialist account of all of morality. He thought that the justification of liberal rights depended only on the theory of distributive justice. If I am correct that his theory of distributive justice was consequentialist, then, had he been successful, he would have successfully completed the consequentialist project for liberal rights.

Unfortunately, the early metaphysical Rawls's theory was flawed. There were two problems with Rawls's theory: (1) the inadequacy of his consequentialist maximin expectation principle as a principle of distributive justice and (2) the failure of his attempt to derive the lexical priority of the Liberty Principle (specifying the relevant autonomy rights) from his consequentialist maximin expectation principle. I discuss the first problem in chapter 4 and the second in chapter 7.

After the failure of Mill to successfully complete the consequentialist project for autonomy rights and the failure of the early metaphysical Rawls to successfully complete the consequentialist project for liberal rights, these consequentialist projects were largely abandoned. The later political Rawls himself disavowed his earlier attempt to give a consequentialist explanation of the lexical priority of the Liberty Principle over the Difference Principle and instead adopted the more promising line of giving it a nonconsequentialist explanation. Over the next 30 years, most of the influential accounts of rights and justice—those of the later Rawls (1993) as well as Barry (1995), Buchanan (2004a), G. A. Cohen (2008), R. Dworkin (2000), Habermas (1990 and 1996), Mills (1997), Nagel (1991), Nozick (1974), Nussbaum (2000), Sen (2009), and Thomson (1990)—were nonconsequentialist.[13]

Are Mill's and the early Rawls's consequentialist projects hopeless? There are many reasons for thinking that they are, especially if the account is

welfarist. The first reason is that such projects seem misguided. Consider, for example, autonomy rights. It seems almost perverse to try to ground a package of autonomy rights on considerations of (appropriately distributed) well-being, when there is a much more direct grounding of autonomy rights in the value of autonomy or consent. Most rights theorists today are some sort of nonconsequentialist, because most of them ground autonomy rights in the value of autonomy or consent, not well-being.

Even if the consequentialist were somehow able to give a consequentialist grounding to the same package of autonomy rights as the nonconsequentialist, the consequentialist explanation would be indirect and complex, whereas the nonconsequentialist account is simple and direct. This seems particularly true for a right against legal paternalism, which was an important element in Mill's package of autonomy rights. On a consequentialist account it would seem that there would be a strong presumption *against* any such right, because, after all, the goal of paternalistic laws is the promotion of well-being. But it seems obvious that a nonconsequentialist account based on the value of autonomy or consent would directly support a right against paternalism.

Thus, anyone who would seek to revive the consequentialist project for autonomy rights or for liberal rights takes on a substantial burden. It is not enough to rig together a Rube Goldberg consequentialist account that just happens to yield the same rights as a nonconsequentialist autonomy-based account. Because the consequentialist account is more complex than the nonconsequentialist account, it must do a better job than the nonconsequentialist account of explaining the contours of the relevant rights—for example, of the contours of an acceptable right against paternalism—than the nonconsequentialist account. The more complex consequentialist account can be favored over the simpler nonconsequentialist account only if the nonconsequentialist account generates explanatory problems that the consequentialist account is able to solve. Of course, the mere existence of explanatory puzzles does not discredit nonconsequentialist accounts, because all accounts have explanatory puzzles. However, it seems to me that there are a number of deep explanatory puzzles for nonconsequentialism that point to a deeper level of explanation, at which level the relevant explanatory principles are consequentialist.

Social Practice Consequentialism as an Explanatory Meta-Theory

In an earlier work (Talbott 2005) I explained my reasons for thinking that moral reasoning is largely bottom-up rather than top-down. Moral reasoning does not begin with principles that are self-evident or rationally intuited. Instead, our moral norms or principles are generally the product of millennia of experience and thought about actual or hypothetical particular cases.

Bottom-up moral reasoning is of two kinds: First, judgments about particular cases can provide support for principles or norms that explain them; second, judgments about particular cases can undermine principles or norms that are incompatible with them.

To explain my consequentialist account of rights, I need to distinguish two levels of moral thought. By *ground-level moral thought*, I mean the moral judgments and moral reasoning involved in a social group's shared practice of moral evaluation. It includes particular moral judgments of rightness and wrongness or justice and injustice ("The system of slavery practiced in the antebellum southern United States was wrong") and norms and principles ("Slavery is wrong"). Though I believe that the discovery of ground-level moral norms and principles is primarily a product of bottom-up reasoning, in a particular situation, ground-level moral reasoning can be either top-down, as when, for example, I conclude that slavery in the United States was wrong because I accept the principle that all human beings have a right not to be enslaved, or bottom-up, as when, for example, I conclude that human beings have a right not to be enslaved on the basis of studying the various institutions of slavery and deciding that each of them is wrong.[14]

Not all moral thought takes place at the ground level. There is another kind of moral thinking that philosophers sometimes do when they theorize about ground-level morality. This is explanatory reasoning, in which the goal is to explain ground-level moral thought. Not all explanations of ground-level moral thought qualify as *moral* explanations, because some explanations of them are debunking explanations. *Debunking explanations* explain ground-level moral thought in a way that implies that it is all a mistake. Thus, for example, a Marxist explanation of ground-level moral thought as a tool to promote the interests of the ruling class or an evolutionary explanation of ground-level moral thought as a "collective illusion of the human race, fashioned and maintained by natural selection" (Ruse 1995, 235) would be a debunking explanation.

In contrast, if an explanation of ground-level moral phenomena is not a debunking explanation—that is, if it at least leaves it open that some of the ground-level moral judgments are true (or morally appropriate)—I refer to it as a *moral meta-theory* and to the principles that it employs as *meta-theoretical moral principles* or *meta-level principles*. Meta-level moral principles are explanatory principles that attempt to explain the moral appropriateness of ground-level moral thought in a way that does not debunk it.

It is important to distinguish between ground-level consequentialism (*direct* consequentialism) and meta-level consequentialism (*indirect* consequentialism). Direct consequentialism has been pretty thoroughly discredited. There are many generally accepted particular moral judgments that conflict with almost any direct consequentialist principles (Nozick 1974, 28). Scanlon and Darwall have reinforced this objection to direct consequentialism by arguing that the concept of well-being itself plays almost no role in first-person moral reasoning (Scanlon 1998, 126–133) and by arguing that desirability (or

good consequences) is a reason *of the wrong kind* to warrant our second-person practice of making moral claims on others or of holding them accountable (Darwall 2006, 15, 104, 192, 311).

Both Scanlon and Darwall seem to take their arguments as arguments against moral consequentialism, but I think this is a mistake. The reason is that they do not seriously consider the kind of *indirect consequentialism* that uses consequentialist meta-principles to *explain* the moral appropriateness of ground-level moral thought (when it is morally appropriate), whether first- or second-person. To refute indirect, meta-level consequentialism, it is not enough to show that ground-level moral reasoning is not consequentialist. It is necessary to consider whether there is a consequentialist meta-theory that explains the moral appropriateness of the nonconsequentialist ground-level moral reasoning.[15]

For example, Brandt (1992) believed that ground-level moral reasoning should be guided by simple *nonconsequentialist* moral rules (e.g., that lying is wrong), but he thought that the meta-principle that *explained* why ground-level moral reasoning should be guided by such rules was a consequentialist (rule utilitarian) one that favored systems of rules that maximized utility. Brandt's account is an indirect consequentialist account, because it uses consequentialist meta-principles to explain the moral appropriateness of nonconsequentialist first-order moral principles.

Similarly, Mill's [1859] theory of robust, inalienable liberty rights was also a consequentialist meta-theory, as any plausible consequentialist account of *robust* rights would seem to have to be. Mill's brand of indirect consequentialism was more general than Brandt's, because it was not limited to explaining the justification of systems of rules, but could be applied to explain the justification of any social institution or practice (e.g., the family), whether or not it could be defined by a system of rules. There is no rule book for being a good parent, nor is it plausible that there could be one. But Mill's consequentialist meta-theory could easily be used to explain the justification of the family as a social practice.[16] I refer to this kind of indirect consequentialism as *social practice consequentialism*.

Mill proposed a consequentialist meta-theory for all ground-level moral thought. As I interpret him, the early metaphysical Rawls (1955 and 1971) had a consequentialist meta-theory, but the meta-theory addressed only distributive justice, not all of morality.

It is an interesting question whether the early Rawls himself thought of the maximin expectation principle as a moral meta-principle or as itself a part of ground-level thought about justice. However, there is no doubt that Rawls thought that the constitutional constraints on legislators and judges would not be consequentialist, but would be given by principles establishing the lexical priority of the rights covered by the first principle of justice, autonomy rights and democratic rights. In a legal system in which judges applied the maximin expectation principle in their decisions, judges would make exceptions to laws whenever they thought it would maximize the expectation

of the least advantaged group to do so. This is not Rawls's view of the role of judges (1971, 196–201).

If Mill proposed a consequentialist meta-theory of all of morality and the early Rawls's theory can be taken to be a consequentialist meta-theory of distributive justice, then my consequentialist meta-theory is somewhere in between the two. In this volume, I attempt to articulate a consequentialist meta-theoretical principle that explains not all of ground-level morality, but only a part of it. But that part turns out to include all of what Rawls thought of as distributive justice. So my consequentialist meta-theory is narrower than Mill's, but broader than the early Rawls's. However, the list of rights that I defend is more expansive than Mill's list of autonomy rights and more expansive than Rawls's list of liberal rights. I refer to them as *human rights* because they are the robust, inalienable rights that all governments should guarantee to all their citizens. Because my consequentialist principle explains the content of those human rights norms, I think of it as providing a consequentialist explanation of human rights.

Primary and Secondary Ground-Level Moral Thought

To classify my kind of meta-level consequentialism, it is necessary to say something more about ground-level morality. For my purposes, it is useful to divide ground-level moral thought into two parts: primary moral judgments and secondary moral judgments. Examples of *primary* moral judgments, norms, and principles are ordinary judgments about the rightness or wrongness of particular actions or kinds of actions—for example, that murder is wrong. The *secondary* moral judgments are moral judgments about the enforceability of other moral judgments—for example, judgments of the permissibility of self- and other-defense, deterrence, and punishment.[17] The judgment that murderers may be imprisoned is a secondary moral judgment. There is an infinite hierarchy of secondary moral judgments. At the first level are judgments about the enforceability of primary moral judgments (e.g., that murderers may be imprisoned). At the next level are judgments about the enforceability of secondary moral judgments on the enforceability of primary moral judgments (e.g., the judgment that a convicted murderer may be punished for attempting to escape from prison). There is no theoretical limit to the number of levels of secondary moral judgments (e.g., the judgment that a person imprisoned for attempting to escape from prison should be further punished for further attempts to escape), but in practice, the number is quite limited.

Because Mill attempted a meta-level consequentialist explanation of all of morality, he assumed the burden of providing a consequentialist explanation of both primary and secondary ground-level moral thought. I am sympathetic to this project, but it is much too large a project for me to take on here. In this book, I limit my explanatory project to primary ground-level

moral judgments. And even here, my project is limited. Although I am sympathetic to the project of providing a meta-level consequentialist explanation of all of ground-level primary moral judgment, even that project is too large for me to take on here. My more modest project is to explain the moral appropriateness of certain improvements to primary ground-level moral thought.

Let me explain. If they persist long enough, all moral traditions change over time. I believe that, at some point in the history of any moral tradition, the moral appropriateness of at least some changes in its primary moral judgments (when they are appropriate) is explained by a consequentialist meta-principle. When a moral tradition has passed this point in its history, I will say that it has crossed the *consequentialist threshold*. Once a moral tradition has crossed the consequentialist threshold, the moral appropriateness of most changes in its ground-level primary moral judgments is explained by a consequentialist meta-principle, which provides a sufficient condition for moral improvement. Because this consequentialist meta-principle turns out to be the main meta-principle in the explanation of the moral appropriateness of human rights norms, I refer to it as the *main consequentialist meta-principle*, or the *main principle* for short.[18]

I have no way of determining exactly when a moral tradition crosses this consequentialist threshold, but every major religious and moral tradition has crossed it. One positive test for whether or not a tradition has crossed this threshold is whether it endorses some version of the Golden Rule.[19] Every major religious and moral tradition has done so.[20]

Why does the main principle come into play only after a moral tradition has crossed the consequentialist threshold? The guiding idea is this: Initially, moral practices are favored in processes of biological and cultural selection for their advantages. We now know that what seem to be moral or proto-moral practices have even been selected for in nonhuman species (e.g., de Waal 2006). At these early levels of moral development, it may be that evolutionary constraints set the standards for improvement in ground-level moral practices. But when a culture reaches a certain level of moral development, those who receive moral training in it acquire a kind of moral sensitivity that replaces imperatives of biological and cultural selection in influencing changes in ground-level primary moral practices. I say something more about this moral sensitivity in chapter 5. There is no way to tell precisely when this transformation occurs, but when a cultural tradition adopts a version of the Golden Rule, we can know that it has occurred. Although the Golden Rule itself is not a rule of reciprocity (as it would be if it enjoined us to love our friends and hate our enemies), when it is adopted as part of a shared moral practice within a social group, it functions to establish mutually beneficial reciprocity relations. Individuals do not usually benefit directly from acting on the Golden Rule, but everyone benefits from other people's willingness to act on it. The main principle is a principle for making moral improvements in a system of moral reciprocity relations.

What about secondary moral practices—those that have to do with enforcement (e.g., norms of punishment)? It is sometimes claimed that there is a version of the Golden Rule that justifies retributive punishment of those who violate primary moral norms (e.g., an eye for an eye, a tooth for a tooth). Call this the *Retributivist Golden Rule*. The Retributivist Golden Rule seems clearly mistaken. Even setting aside its morally problematic implications (e.g., that the proper punishment for rapists is to be raped), it is clearly inadequate as a secondary moral principle. Consider the crime of stealing $1,000. It is easy to see that the appropriate punishment for that crime may be much more than restitution and a fine of $1,000, because such a fine would not effectively deter stealing if the probability of getting caught was less than one-half. Thieves would make money if they just regarded having to pay the fine when they got caught as one of the costs of doing business.

Even though the Golden Rule does not seem to apply to secondary moral thought about punishment of the guilty, it seems to me quite plausible that secondary moral thought can be explained by a consequentialist meta-principle. However, to try to carry out the explanation is beyond the scope of this book. For present purposes, it is necessary to narrow the focus to the project of providing a meta-level consequentialist explanation of improvements in ground-level primary moral thought, for moral traditions that have crossed the consequentialist threshold.

Rawls was able to make a clean division between primary and secondary moral thought by assuming strict compliance with his two principles of justice (1971, 8). If there is strict compliance, enforcement is unnecessary, and so a meta-level theory of the enforceability of moral judgments is also unnecessary. However, I do not adopt Rawls's extremely idealized assumption of strict compliance. I discuss the justification of moral norms and principles, and, especially, human rights, in more realistic cases in which it is known that enforcement will be necessary. In such cases, it is often thought that consequentialists must allow for legal systems that knowingly punish the innocent (Nozick 1974). In chapter 6, I argue that everyone, consequentialist or nonconsequentialist, must allow for legal systems that are known to punish some innocent defendants (because it is inevitable that some innocent defendants will be mistakenly judged to be guilty), though neither is committed to endorsing a system that punishes defendants known to be innocent.

My main focus will be on ground-level moral thought about human rights and, by extension, the constitutional provisions or laws that guarantee them. Included in ground-level moral thought are the rationales that judges give to justify their legal decisions, when those decisions overrule prior law or apply old law to a new kind of case. Also, included are the rationales that legislators give to justify constitutional amendments or to justify laws, when the considerations are considerations of justice, rather than considerations of how best to promote the interests of their constituents. When judges' or legislators' rationales involve considerations of justice or fairness, they are a part of ground-level moral practices that the main principle applies to.[21] It is

important to keep in mind that when I say that the main principle applies to changes in ground-level moral thought, I mean to include this kind of legal thought, also, because, in this way, the main principle is the most important principle for explaining the appropriateness of changes in human rights. The main principle does not apply to secondary norms, so, in all the examples I discuss, I just assume that the enforcement provisions of the relevant laws satisfy the relevant proportionality constraints.

Given these preliminaries, I can simplify my exposition by assuming, unless I say otherwise, that by *ground-level moral and legal thought* I mean changes in ground-level primary moral and legal thought in a moral tradition that has passed the consequentialist threshold. That is the ground-level moral and legal thought that my consequentialist meta-theory aims to explain the moral appropriateness of.

My Explanatory Strategy

Consequentialists typically begin by defining the important terms (e.g., *well-being*) and then offering some formula for rightness or justice in terms of well-being (e.g., in terms of maximizing overall well-being). I don't have direct rational insight into self-evident truths about morality and justice, so I can't define any of the most important terms that I use and I cannot provide a precise formula for rightness or justice.

Particular Moral Judgments

My approach is to work primarily in the other direction, bottom-up rather than top-down. As I see it, ground-level moral principles (including principles of human rights) are the result of a largely bottom-up process of discovery, based on ground-level *particular moral judgments*—that is, moral judgments about actual and hypothetical particular cases (e.g., that Hitler's attempt to exterminate the Jews was wrong). It is important to understand how we can discuss particular cases. Here is an example: Typically, in discussing particular cases, I assume that an increase in life expectancy represents an increase in well-being. In such cases, I am depending on your ability to imagine cases in which it does increase well-being, because I know that, for any interesting generalizations that I might formulate, there will almost always be exceptions. It is not always true that increases in life expectancy increase well-being. For example, we can easily imagine cases in which someone with a fatal disease faces a short period of suffering that will end with death. Prolonging their period of suffering would not generally be a way of promoting well-being. Notice that, even here, I am relying on your ability to imagine cases of the relevant kind, because I am not denying that there are exceptions to the exception—that is, cases in which it would promote well-being to extend the period of suffering—for example, if living for 2 days

more would allow time for to reconcile with an estranged family member, something that was very important to the suffering person, before dying.

How are we able to refer to the relevant kinds of examples with finite descriptions, if adding further information can change our moral judgment about a particular case? This is a fascinating question that deserves more attention than I can give it here. Part of the answer is that we do it by making our intent clear to our audience (e.g., the intent to describe an example of an increase in well-being) and then providing the audience with enough information for them to be able to imagine the relevant kind of example (e.g., making available a drug that significantly increases life expectancy). When my audience knows that I intend to describe an example involving an increase of well-being and then I ask the audience to imagine a case in which a certain drug increases life expectancy, the audience will look for examples of drugs that increase life expectancy in a way that increases well-being. If such examples are difficult for the audience to find, then I did not provide enough information. But if the information that I provided makes such examples easy for the audience to find, then there is no need to provide more information. Providing more information might rule out some exceptions, but there is no need to think that we must be able to describe examples in a way that rules out all exceptions, in order to be able to discuss particular cases.[22]

Ground-Level Moral Norms and Principles

How can we explain particular moral judgments? The simplest kind of explanation would be an explanation in terms of ground-level moral generalizations—that is, ground-level norms or principles. For example, the norm "coercion is wrong" would explain the variety of particular cases involving wrongful coercion. But what about cases, hypothetical as well as actual, in which coercion is not wrong? J. S. Mill gives the example of a person about to cross an unsafe bridge ([1859], 109). If there is not time to explain the danger, the use of force is permitted to stop that person from crossing the bridge.

Typically, the first reaction to the discovery of exceptions to ground-level moral norms is to try to fix the norm by building exceptions into it or by finding a more general ground-level principle that explains why the norm holds in those cases in which it does and why it fails to hold in the exceptional cases. For example, one might propose a new ground-level norm: Coercion is wrong unless necessary to prevent death or serious injury in a case in which there is no time to explain why there is a danger of death or serious injury. Or one might instead seek a more general ground-level principle, and hit upon a version of the Golden Rule: Do unto others as you would have them do unto you. This version of the Golden Rule would prohibit most cases of coercion, but would allow an exception in the case of the unsafe bridge.[23]

One of the great puzzles of moral philosophy is that this process of adding exceptions to our ground-level norms or finding new ground-level norms to cover the exceptional cases so far discovered does not ever seem to end. There

are always more exceptions to the ground-level norms or principles. This result is so familiar that Scanlon simply assumes that the moral principles he discusses are actually "labels for much more complex ideas" (1998, 199) that cannot be captured in a simple rule. Because of the potential for exceptions to a given principle, exceptions to the exceptions, and so forth, Scanlon thinks there must be an indefinite number of moral principles (201). Dworkin makes the same point about principles in the law: They all have exceptions (1977, 25).

But if ground-level moral and legal norms and principles typically have exceptions, there is no complete explanation of a particular moral judgment at the ground level. The reason is simple. If by "coercion is wrong" we understand "coercion is usually wrong," then the norm cannot by itself explain the wrongness of a particular act of coercion, because the full explanation would require not only the norm that coercion is usually wrong but also an explanation of why the relevant particular case is one of the "usual" rather than the "unusual" cases.

Meta-Level Moral Principles

Of course, it may be that there are true exceptionless ground-level norms or principles that explain all the true ground-level particular moral judgments and our problem is just that we have not yet discovered them. The alternative that I want to seriously consider is that there is a higher level of explanation at which it is possible to explain the moral appropriateness of ground-level moral judgments, including particular moral judgments, norms, and principles (when they are appropriate) and their moral inappropriateness (when they are inappropriate).[24] Surprisingly, at the meta-level, we discover an exceptionless principle that not only explains the moral appropriateness of changes to the ground-level particular judgments, norms, and principles, but, as I show in chapter 5, it also explains why substantive ground-level norms and principles always (or almost always) have exceptions. I call this meta-level consequentialist moral principle the *main principle*. The main principle explains the moral appropriateness of most *changes* in ground-level primary particular moral judgments and ground-level primary norms and principles (when they are morally appropriate) in moral traditions that have passed the consequentialist threshold.

Puzzles about What Is to Be Explained

I have said that the main principle explains the moral appropriateness of most changes in ground-level moral and legal thought (when they are morally appropriate). For ease of exposition, let's focus on moral thought. The application to legal thought is exactly parallel. When we ask how we can test such a theory, a deep puzzle emerges. One way to test the theory would be to look back on the history of ground-level moral thought to identify the cases

in which exceptions have been made to accepted ground-level norms or principles. But this would be a fallacious test, because a moral meta-theory is not a descriptive theory. It is not an attempt to explain all of the changes in ground-level moral thought that have actually occurred. It is an attempt to explain the moral truth or appropriateness of those changes that were morally true or appropriate (and moral falsity or inappropriateness of those that were morally false or inappropriate).

So it seems that we must test the theory against our own considered judgments about which of the changes were morally appropriate—or, to be more exact, which were moral improvements—and which were not. This is a cause for worry. What is to keep me from adjusting my judgments about which moral changes have been improvements to fit my theory?

In addition, it would seem that any such a theory would be hopelessly relativistic, because there is so much disagreement about which moral changes have constituted improvements. I regard the extension of equal rights to women as an important moral advance, but the Taliban regards it as an example of moral degradation. Who is to decide which changes qualify as improvements? In this book, I do not maintain neutrality between different views of moral progress.[25] But the test of my theory is not that it persuades me or people who share my beliefs about moral progress. As I explain in chapter 7, the best test of both my theory and the Taliban's theory (though not an infallible one) is how they fare in the process of free give-and-take of opinion. Of course, if the Taliban had their way, they would suppress the process of the free give-and-take of opinion. They could thus prevent any challenges to their claim that their theory was justified. But claiming it would not make it so. I discuss these issues more fully in chapters 7 and 14.

There is another problem, also. Any adequate theory of moral improvement must have implications that go beyond the actual changes that have occurred in the past. It will have implications for which potential future changes would be moral improvements and which would not. Are we supposed to test these implications against our *current* judgments of which future changes would be moral improvements and which would not?

That cannot be a satisfactory test. When we look at the past history of changes that we now regard as moral improvements, we find that there were times when most people had a moral blind spot that prevented them from recognizing that the change would be a moral improvement. For example, very few Europeans raised moral objections to the slave trade in the sixteenth century. Even in the eighteenth century, the slave trade flourished and some of the authors of the Declaration of Independence were able to hold that all men are created equal while also defending slavery. Given the prevalence of moral blind spots in the past, it would be a display of *hubris* to think that we ourselves don't also have moral blind spots. But if we have moral blind spots, then there are some changes to our own ground-level moral thought that would be moral improvements, but, due to our own moral blind spots, we don't realize that they would be.

This has the following paradoxical implication: Suppose someone articulated a moral meta-theory that compellingly explained the moral appropriateness of past changes in ground-level moral thought now regarded as moral improvements and also identified exactly those changes in current ground-level moral thought that would now be regarded as moral improvements. Such a theory might be very useful for many purposes, but we would be almost certain that it was false, because it would fail to identify our own moral blind spots. A fully adequate moral meta-theory must identify some potential improvements in ground-level moral thought that we would not today regard as improvements.

Thus, there is no adequate synchronic test of a moral meta-theory. A moral meta-theory must be tested, in part, diachronically, by the way that ground-level moral thought changes in the future. Because future changes in ground-level moral thought can themselves be influenced by our moral meta-theories, a moral meta-theory theory can be tested not only by its predictions about changes that will in the future be regarded as moral improvements, but also by the changes in ground-level moral thought that it contributes to bringing about.

Another way of putting this point is to say that a moral meta-theory is a theory of past changes in ground-level moral thought that have been improvements and of potential future changes that would be improvements. The puzzle is that we must test such theories by our own ground-level moral thought, which we have good reason to believe is itself subject to improvement. Of course, if our ground-level moral thought is massively mistaken, our moral meta-theories will be massively mistaken also. However, I do not mean to be raising skeptical worries here.[26] A moral meta-theory that provided a satisfactory explanation of the moral appropriateness of past changes in ground-level moral thought that we now take to have been improvements and a satisfactory explanation of the moral appropriateness of potential future changes that future generations will come to regard as improvements would be a stunning accomplishment. That would not only be good evidence for its truth, it would be the best possible evidence for its truth.

Improvements Are Comparatively Better, Not Optimal

Because my goal is to explain moral improvements, my consequentialist theory is not an optimizing theory. It is almost certain that no human society will ever discover an optimal moral system, on any reasonable criterion of optimality (cf. Sen 2009). But all human societies can improve their moral practices. The principle that explains which changes are improvements has to make comparative evaluations of only a relatively small number of relevant alternatives—usually, the *status quo* and one or two potential changes to the *status quo*. It is much more likely that human societies could satisfy such a comparative principle than that they could ever satisfy any plausible optimizing principle.

Strict Universality of Particular Moral Judgments and of Meta-Level Principles

One of the most surprising claims in my first volume (Talbott 2005) was that by reasoning in a largely bottom-up manner it is possible to discover fundamental moral principles that are strictly universal—that is, true of all rational beings in all possible worlds. Traditionally, it was thought that the only way to have knowledge of strictly universal (i.e., metaphysically necessary) principles of any kind was through direct *a priori* insight. Because I don't claim to have any direct *a priori* insight, it is surprising that I would claim that there are strictly universal moral meta-principles and that we are engaged in an ongoing historical process of trying to figure out what they are.

The key to understanding how it would be possible to discover such principles is to understand that our true particular moral judgments are also strictly universal, though in a slightly different sense. When I make a particular moral judgment (e.g., that it was wrong of the Western European colonists to enslave American natives or to force them to adopt the Christian religion) I do not claim to be infallible. But I do think that we are justified in placing a great deal of confidence in them in clear cases. These judgments are largely true, and when they are true, they are objectively true. They are true not just for human beings or for those who share our moral tradition. When they are true, they are true for any rational being. This is the sense in which particular moral judgments can be strictly universal. If this is right, then we can use particular judgments about actual and hypothetical cases to support principles that apply to actual and hypothetical cases. Were we to discover the fundamental principles that explain all actual and hypothetical cases, they would be true in all possible worlds. So they would be strictly universal principles.

For most of human history, the goal of moral inquiry has been to formulate exceptionless ground-level moral principles. The failure to do so has led many philosophers to deny that there are any (Dancy 2004). It is somewhat surprising that there might be a meta-level explanation of why exceptionless ground-level moral principles are so rare and even more surprising that the meta-level explanation would employ an exceptionless meta-theoretical principle. But there is and it does, as I explain in chapter 5. So it turns out that there are strictly universal moral principles, but they are meta-theoretical principles, not ground-level principles.

Contingent Universality of Human Rights Norms or Principles

Because human rights norms or principles are ground-level norms or principles, we should not expect them to be exceptionless. I have already acknowledged that they are not, when I said that the project is to explain human rights that are *robust* but not *absolute*. Some readers will be disappointed by

this admission. They will not want to give up on the search for exceptionless ground-level human rights principles. I hope that the explanation of why it is almost inevitable that ground-level principles have exceptions in chapter 5 will help to reconcile those readers to this result.

If ground-level human rights principles are not exceptionless, then they are not strictly universal, not true in all possible worlds. This should not be surprising. Human rights depend on human capabilities. In possible worlds in which human beings had very different capabilities, they would be expected to have very different rights.

The universality of human rights is not strict universality, but it is an important kind of contingent universality: Given what we know about human beings and human societies in this world, the main principle explains why it is morally appropriate that *all* human societies guarantee autonomy rights (and other rights on my list of human rights) for *all* normal adults. This is the sense in which the rights on my list of human rights *should be universal*.

Justifying Government Coercion

Although the main principle applies to exceptions to any ground-level primary moral norm or principle, a particularly important category of exceptions is the category of exceptions to the ground-level moral prohibition on coercion, because that is the prohibition to which judges and legislators must be able to justify exceptions if they are to be able to justify making new law and modifying old law. Laws are coercive. Because coercion is generally wrong, the rationale for coercive laws must state an exception to the general rule against coercion. To a first approximation, the main principle supports exceptions to the general rule against coercion when such exceptions, evaluated as a social practice, equitably promote the well-being of those who are coerced. There are two important kinds of laws that can be used to promote well-being:

(1) Paternalistic laws. These are laws that limit a person's liberty for her own good, even though the person herself may disagree. In chapters 12 and 13 I argue that when certain basic rights are guaranteed, normal human adults should have liberty rights to freedom from government paternalism, unless it satisfies a special kind of hypothetical consent standard, the most reliable judgment standard. The most reliable judgment standard is a ground-level standard. The explanation of its moral appropriateness is a meta-theoretical explanation that employs the consequentialist main principle. So the standard for rights against paternalism that I articulate in chapter 12 is not consequentialist, but the explanation of why that standard is morally appropriate is.

(2) Legal solutions to collective action problems (CAPs).[27] This is the most important category of laws promoting well-being. CAPs are ubiquitous. Climate change, pollution, and fisheries destruction are negative examples,

in which the outcome is bad if everyone drives gas guzzlers, pollutes, or overfishes, but in which an individual person's contribution to the badness of the outcome is so small as to be negligible and each individual has a reason to do the slightly bad act, because it is more costly to her not to do it. Fire protection, highways, traffic signals, and medical research are all positive examples, in which the outcome is good if everyone contributes, but each individual's contribution to the good outcome is so small as to be negligible and each individual has a reason to avoid doing the slightly good act because it is more costly to her to do it. I refer to these problems as *N-Person Prisoners' Dilemmas*.[28] In such situations, by convention, the act that leads to the better results when chosen by everyone is called *cooperating*. The act that leads to the worse results when chosen by everyone is called *defecting*. A quick test for an N-Person PD is whether there would be some temptation to free ride—that is, to defect if everyone else or almost everyone else is cooperating. This test shows that typical cases of stealing, cheating, lying, promise-breaking, even murder, also generate an N-Person PD. Those who steal benefit from others' not stealing to be able to enjoy the benefits of their theft.

CAPs are not only ubiquitous, but they are what might be called *productive*: A solution to one can and often does generate others. Market economies are a solution to a CAP (the productive investment CAP), but they themselves generate possibilities for corporate fraud and market bubbles, both CAPs. Governments are a solution to CAPs, but voting itself is a CAP. Not all CAPs should be solved. The main principle explains why. Price fixing is a solution to a producers' CAP that the main principle does not endorse.

Human societies could not thrive unless they had ways of solving CAPs. Indeed, human societies would probably not exist were it not for CAPs. It is the existence of CAPs that gives an evolutionary advantage to social species, such as human beings (Wright 2000). Recognizing and sanctioning cheaters and other free riders is so important to a social group that evolution has almost surely endowed us with the psychological equipment to detect cheaters and respond appropriately to them (Cosmides and Tooby, 1992). Morality itself is one social practice that helps to solve CAPs. A legal system is another.

Hobbes [1651] thought that life without a government to make and enforce laws would be so awful that any government, no matter how bad, would be infinitely better than no government at all. That is an exaggeration, but it is not an exaggeration to say that solving CAPs is the most important function of a government. Governments implement coercive solutions to CAPs by punishing defectors. They imprison murderers and thieves, fine polluters, establish fire departments, fund medical research, establish and administer a system of police, courts, and prisons to enforce their laws, and punish those who don't pay their taxes to pay for all these solutions to CAPs.

In this book I argue that, when established against a background of the other basic rights, part of the rationale for constitutionally limited democratic rights is their role in solving CAPs in a way that equitably promotes

well-being. The ground-level principles establishing democratic rights are nonconsequentialist (e.g., one person, one vote). It is at the meta-level, where the rationale for ground-level democratic principles is consequentialist—that establishing such rights is a good way of solving CAPs in a way that equitably promotes well-being (at least, better than the other relevant alternatives). As I explain in chapter 10, the main principle endorses democratic rights, in combination with constitutional protections for robust, inalienable human rights, because of their tendency to equitably promote well-being.

What Is Normative Truth?

I believe that the main principle is a *true* meta-level moral principle. To say this is to say that some changes in ground-level moral practices *really are* improvements. They are not just improvements from a liberal point of view or a cultural or religious point of view or a species point of view. They are improvements from the objective point of view (Nagel 1986).

To believe in normative truth, it is not necessary to believe in any weird entities or forces. All that is required is to believe that there can be *real* moral progress. If there are true moral meta-principles, they are not written on stone tablets. What kind of truths are they? In a sense, this entire book is an extended answer to that question.

Moral truths are only one category of normative truth. There are also normative truths about what it is rational to believe and truths about what it is rational to do in nonmoral situations. In all these cases, it is very difficult to articulate exceptionless principles, but not so difficult to describe some clear examples of rational and irrational belief or rational and irrational action (in nonmoral contexts) or moral or immoral action. In all three of these cases, if there are truths, they are not purely descriptive truths about what people actually do or believe, but truths about what any rational or moral being should or should not do or believe. Anyone who believes in normative truth is a *normative realist*.

Some people think that there could be normative truths only if God made them true (e.g., E. O. Wilson 1998, chap. 11). These people must think there are no objective normative constraints on what God believes or does. This is a puzzling view. Could God have made it true that it was rational to believe all contradictions—for example, that God exists and does not exist or that God is omnipotent and not omnipotent? Could God have made it true that it is morally right to torture innocent children merely for fun? These questions are puzzling enough to motivate our taking seriously the possibility that normative truths are constraints on all rational beings, including God.

Some people are moral noncognitivists. They think there aren't any normative truths (e.g., Gibbard 1990). Noncognitivists think that our normative avowals evince a certain kind of attitude. There is nothing objective for those avowals to correspond to or to fail to correspond to. These views are close

relatives of the views of those who think that God makes normative truths. If there is no God to make them *true (or false)*, then it is human attitudes and emotions that make them *appropriate (or inappropriate)*. One way to be a normative realist is to believe that human attitudes could not make the law of noncontradiction appropriate or inappropriate and could not make torturing children merely for fun appropriate or inappropriate.[29]

But really, I can hear someone say, isn't it enough to find normative principles that apply to all human beings? Why think there are any strictly universal normative truths that apply to all rational beings? To answer these questions, it is useful to consider the example of utilitarianism.

What Universal Moral Truths Might Be Like

Though the utilitarian principle of maximizing overall (i.e., total or average) well-being is not an adequate principle of morality, either as a ground-level or a meta-level principle, it is close enough to give us some idea of how the bottom-up process of moral inquiry might lead us to a strictly universal moral principle. It really does seem that well-being is something that would be important to any rational being. The problem with utilitarianism was that it mistakenly concluded that maximizing overall well-being would be important to any rational being. As Rawls (1971) pointed out, institutions that maximize well-being need not promote everyone's well-being. Thus, utilitarianism allows for the possibility of reducing some people's well-being in order to produce a more-than-offsetting increase other people's well-being. This is utilitarianism's distributional blind spot. The main principle does not have a distributional blind spot. It aims at equitably promoting everyone's well-being. Distribution matters.[30]

Why does the main principle apply to all rational beings? Consider only one application of the main principle: determining morally appropriate norms for solving CAPs. CAPs are practically unavoidable for rational beings who interact with other rational beings.[31] And even rational beings who never find themselves in a collective action problem with other rational beings could still ask themselves what they should do if they ever were to find themselves in such a situation. So we should at least entertain the possibility that there might be principles that determine the moral appropriateness of norms for solving CAPs for any kind of being. If so, they would be strictly universal moral meta-principles. The possibility of such principles should not be ruled out at the beginning of our inquiry.

The Main Principle and Human Rights

In the previous volume, I outlined nine basic human rights. I called them *human* rights, because they are rights that should be guaranteed to all normal

human beings and *basic* because they must be guaranteed for a government to meet a minimum standard of moral legitimacy.[32] Here is the list of the basic human rights:

1. A right to physical security
2. A right to physical subsistence (understood as a right to an opportunity to earn a subsistence for those who are able to do so and a welfare right for those who are not)
3. Children's rights to what is necessary for normal physical, cognitive, emotional, and behavioral development, including the development of empathic understanding
4. A right to an education, including a moral education aimed at further development and use of empathic understanding
5. A right to freedom of the press
6. A right to freedom of thought and expression
7. A right to freedom of association
8. A right to a sphere of personal autonomy free from paternalistic interference
9. Political rights, including democratic rights and an independent judiciary to enforce the entire package of rights

It is useful to group these rights into a small number of partially overlapping categories. I refer to the first eight items on the list as *autonomy rights*, because they are the rights that are necessary for citizens to develop and exercise their autonomy.[33]

The eight autonomy rights can be further divided into *development-of-judgment* rights (the first four rights on the list), because they are necessary to develop the capacity for good judgment (the ability to make reliable judgments about one's own good) and *exercise-of-judgment rights* (the next four rights on the list), because they are necessary for someone who has the capacity for good judgment to actually have good judgment and exercise it. I have more to say about some of these rights in coming chapters.

The final item on the list, political rights, is necessary to make governments appropriately responsive to the judgments of their citizens. An independent judiciary is necessary to protect all of the items on the list from government abuse or majority tyranny.

In the previous volume, I discussed both consequentialist and nonconsequentialist rationales for the nine basic rights, without choosing between them. In this volume I choose. I believe that the consequentialist main principle is the best meta-level explanation of why governments should guarantee their citizens the nine rights on my list. For a government to reliably promote the (appropriately distributed) well-being of its citizens, it must guarantee the nine basic rights on my list.

In this volume, I discuss some of the basic human rights in more depth—security rights (chapter 6); a right to freedom of thought and expression and the

related right of freedom of the press (chapters 7 and 8); democratic rights (chapter 10); and a liberty right against legal paternalism (chapters 12–13). In addition, I identify five further kinds of human rights—that is, robust, inalienable rights that should be universal:

10. Economic rights (chapter 9)
11. Negative opportunity rights—that is, rights against discrimination (chapter 11)
12. Positive opportunity rights—rights to certain capabilities (chapter 11)
13. Social insurance rights (chapter 11)
14. Privacy rights (chapter 13)

Is my list of human rights too long? It would be too long if my goal were to identify the rights on which there currently exists an international overlapping consensus. However, whatever international consensus exists today leaves lots of room for improvement. My list is intended to point to what the consensus *should be*, and to explain why.

How to Make a Case for Consequentialism at the Meta-Level

In the competition between consequentialist and nonconsequentialist theories of human rights, the consequentialist starts out at a distinct disadvantage. The ground-level human rights principles themselves are nonconsequentialist, so they seem to invite a nonconsequentialist explanation. In addition, one of the most important categories of human rights on my list is the category of autonomy rights, and it seems almost self-evident that a nonconsequentialist explanation of autonomy rights in terms of the importance of autonomy would be a simpler and more direct explanation than a consequentialist explanation in terms of equitably promoting well-being.

It turns out that the simplicity of the nonconsequentialist account is also its Achilles heel. To compare the two accounts, we must consider not only how directly and simply they explain the relevant categories of rights, but also how well they are able to explain the contours of the rights in each of the categories. Because the contours of the individual rights involve many nuances and irregularities, a simple theory has difficulty in adequately explaining them. I will try to show that a consequentialist account does a much better job.

At the most fundamental level of analysis in a theory, one finds the central concepts of the theory. In nonconsequentialist theory, two of these fundamental concepts are autonomy and consent. In this book, I try to show that there is an even more fundamental, consequentialist, meta-theoretical level of explanation at which level we can explain the moral significance of these concepts. I discuss the significance of consent in chapter 9 and the nature and significance of autonomy in chapters 12 and 13.

One way of trying to cast doubt on nonconsequentialist theories of human rights is to raise puzzles for them, especially puzzles that seem to have a consequentialist solution. So I raise lots of puzzles in this book. For example, in chapter 6, I show how, in theory, it could be a moral improvement to do away with punishment altogether and how, in theory, a move to a system of strict criminal liability could be a moral improvement. In chapter 8, I show that Rawls's and Habermas's theories fail to support a constitutional right to freedom of expression that includes the expression of intolerant subversive advocacy. In chapter 10, I show how, in theory, we could be warranted in replacing democratic elections with a system of deliberative polling, and I provide a consequentialist solution to the puzzle of why any rights should be inalienable. In chapter 11, I raise a puzzle for views, such as Dworkin's (2000), that are based on the distinction between brute luck and option luck.

In the book, I give extended critical consideration to many of the most influential nonconsequentialist theories, including those of Nozick (in chapters 2–3), Thomson (in chapter 4), Rawls (in chapters 7 and 10), Habermas (in chapters 7 and 10), Dworkin (in chapter 11), and Feinberg (in chapter 12) and briefer critical consideration of many others.

It is important not to overstate the significance of the puzzles I raise for nonconsequentialists. As Kuhn (1962) observed about scientific theories, every theory has its puzzles. One way to allay doubts about my consequentialism is to address and resolve some of the well-known puzzles for consequentialism. So I do, including many of the standard objections to theories based on well-being in chapter 4, Nozick's Wilt Chamberlain example in chapter 3, examples of punishing the innocent and organ harvesting in chapter 6, the problem of seeming to justify lots of paternalism in chapters 12 and 13, and a host of objections in chapter 14. But I could never eliminate all puzzles for consequentialism. In philosophy, every theory has its puzzles.

For that reason, in deciding among theories, the decision often comes down to such considerations as the way a theory unifies disparate phenomena, illustrated by the way that my account in chapter 5 provides a unified explanation of the defeasibility of moral and legal reasoning (and, potentially, all reasoning); or the way that my account of the role of tort law in chapter 9 unifies a market economy and a system of tort law into a single self-regulating system; or by the way the main principle in chapter 3 and the Millian epistemology in chapter 7 unify the seemingly disparate rights on the U.N. Universal Declaration of Human Rights, as explained in chapters 6–13.

To many people, it seems obvious that the moral appropriateness of human rights norms could not be based on their contribution to equitably promoting well-being. Even worse, they see it as a threat to the human rights movement to even suggest that it might be. These people realize that the most influential arguments *against* human rights are typically based on well-being—for example, that poor countries can't afford human rights because they need to

encourage economic development. Because considerations of well-being are usually used to argue for exceptions to human rights, many people rightly fear that even to think of human rights as ways of equitably promoting well-being would make them much less secure. I discuss the paradox of direct consequentialism at great length in chapter 5, in part to try to allay this concern, and then, in chapter 14, I respond to this objection directly. Though the main principle explains the exceptions to human rights norms, it does not support our using *it* as a ground-level principle to justify exceptions to human rights norms.

Conclusion

Mill made the first attempt at a meta-level consequentialist explanation of the moral appropriateness of autonomy rights. As I interpret him, the early metaphysical Rawls expanded the project to try to provide a meta-level consequentialist explanation of the moral appropriateness of both autonomy and democratic rights (i.e., liberal) rights. My goal is even more ambitious: to try to provide a meta-level consequentialist explanation of autonomy rights, democratic rights, and five other categories of rights, as well—economic rights, negative and positive opportunity rights, social insurance rights, and privacy rights. The project is to explain why robust, inalienable rights of all those kinds should be universally guaranteed to all normal human adults by governments everywhere—that is, to explain why these rights *should be* universal. For that reason, I think of them as the rights that should be recognized as *human rights*. My ultimate goal is to provide a consequentialist meta-theoretic explanation of the content of these human rights. This is the *consequentialist project for human rights*. No single book could complete the project, so my aspiration for this book is to contribute to the project and, thus, to make it more plausible that the project might be successfully completed.

Because the methodology for my contribution to the consequentialist project for human rights is largely bottom-up, I can undertake the project even though I have no definition of *well-being* and I have no formula for its equitable distribution. A meta-theory of human rights is a theory of a moving target. If such a theory were to provide a complete vindication of current opinions about what is just or about what human rights should be, it would be a failure, because there is nothing more certain in moral matters than that current opinions can be improved.

A meta-theory of human rights should provide guideposts for potential improvements in current opinions and provide the resources for understanding why future changes are improvements (when they are). This is a tall order for any theory to have to fill. And it is one that any theory is bound to come up short on. Any normative theory of justice or of human rights, including this one, is bound to be imperfect, and thus improvable. This potential for

improvements in our current opinions and in our normative theories is the basis for a dynamic between theory and practice in which, over time, a theory can help us to improve our ground-level moral judgments and our ground-level judgments can help us to improve the theory. My goal in this book is to contribute to that process.

It is important at the outset for me to address a potential misunderstanding. Some people think that a right cannot be a human right if there is reasonable disagreement about it. As I explain in chapter 8, I think this seriously misunderstands the historical-social process by which human rights have been and are being discovered. In any case, let me say right here that almost everything in this book is subject to reasonable disagreement.

TWO

Exceptions to Libertarian Natural Rights

In the previous chapter, I proposed that we pay attention to the historical process of making exceptions to ground-level primary moral norms and principles. In this chapter, I compress and idealize some of that history to briefly illustrate the bottom-up reasoning involved, as an example of what it is that the main principle is designed to explain. I have claimed that the main principle provides a sufficient condition for the moral appropriateness of changes in ground-level primary moral judgments in any tradition that has passed the consequentialist threshold. I illustrate this claim by considering the natural rights tradition that developed in the West, because of the great power of that tradition and because the best way of explaining my theory of human rights is as a development from that tradition. However, it is important to realize that the main principle transcends any particular moral tradition to apply to all moral traditions that have passed the consequentialist threshold.

In the following conversation, three philosophers attempt to formulate ground-level moral principles for the state of nature, a situation in which there are no governments and thus no legal obligations. The state of nature is a heuristic for thinking about moral obligation in a way that avoids confusing it with legal obligation. The state of nature can play this heuristic role without our being committed to thinking that any such state ever actually existed.

An Example of Changes in Ground-Level Moral Principles through Bottom-Up Reasoning

Three philosophers, Moses, Fred, and Bob, were discussing the state of nature. Moses asked them to consider the following example: Anne is sitting minding her own business. Adolph comes up to her, pulls out a gun, and threatens to kill her unless she will be his slave.

Moses, Fred, and Bob all agreed that it would be wrong for Adolph to coerce Anne in this way. Moses suggested the following ground-level principle to explain why it would be wrong:

The Simple Prohibition on Coercion. It is wrong to coerce another human being by threatening to kill her.

Fred disagreed with Moses' principle and gave the following counterexample: Suppose that Winston sees Adolph threaten Anne, so he takes out a gun and threatens to kill Adolph unless Adolph stops threatening to kill Anne and leaves her alone.

Moses, Fred, and Bob all agreed that Winston's threat to kill Adolph was not wrong.[1] Because this conclusion was incompatible with the simple prohibition, they gave it up and looked for another ground-level principle. Fred suggested the following:

The Minimization of Coercion Principle. We should act so as to minimize the total amount of coercion.[2]

Fred suggested that in the previous example, Winston's threat was less coercive than Adolph's, because it would prevent Adolph from doing only one thing (coercing Anne), whereas Adolph's coercion would prevent Anne from doing anything that Adolph did not want her to do.

Bob was not satisfied with Fred's minimization of coercion principle as an explanation of the example, because it treated all coercion as morally on a par. Bob thought this was a mistake. Bob said that some coercion is bad (e.g., Adolph's), but that coercion can be good when it prevents bad coercion (e.g., Winston's). The problem with treating all coercion on a par is that it opens up the possibility of allowing some relatively small amount of bad coercion if the only way to prevent it required a relatively larger amount of good coercion. Bob thought this was unacceptable.

Bob argued as follows. Imagine two possible worlds: In W1 there is enough good coercion to eliminate a lot of bad coercion, but bad coercion still exists; in W2 there is a lot more good coercion, enough to completely eliminate all bad coercion. W2 might be a world in which there are penalties for bad coercion that are so effective that no one ever performs the kind of coercion that would trigger the penalties. According to Fred's principle, W2 would not be morally justifiable unless the total amount of coercion in W2 was less than the total amount in W1. But this is not right. Because W1 contains some bad coercion and W2 does not, W2 is morally preferable to W1, regardless of what the total amount of coercion in each world may be.

This kind of example led Bob to look for a different principle that would distinguish good from bad coercion. Moses suggested the following principle:

Qualified Prohibition on Coercion. It is wrong to threaten to kill another human being, unless one is threatening to kill someone to prevent that person from threatening to kill someone else.

Bob did not accept this principle either. To explain why not, he added to the example of Winston, Adolph, and Anne: Suppose Adolph's friend Benito happens by and sees Winston threatening Adolph. After the others explain to Benito what has happened, Benito takes out a gun and threatens to kill Winston unless he stops threatening Adolph. When Winston stops threatening Adolph, Adolph renews his threat to Anne. According to the qualified prohibition with

one exception proposed by Moses, Benito's threat would not be wrong. Moses, Fred, and Bob all agreed that it would be wrong, so they needed another principle.

Bob thought of a new ground-level principle to propose. In order to motivate it, he added to the previous example: Suppose Winston's friend Franklin happens by and sees Benito threatening Winston. After the others explain to Franklin what has happened, Franklin takes out a gun and threatens to kill Benito unless he stops threatening Winston. When he stops threatening Winston, Winston renews his threat to Adolph, which stops Adolph from threatening Anne.

Moses, Fred, and Bob all agreed that it would not be wrong for Franklin to threaten Benito. They now recognized a pattern in the examples. Bob pointed out that they could identify different levels of coercion. Adolph illustrated level-1 coercion, coercion of someone who was not coercing anyone else. Winston illustrated level-2 coercion, coercion aimed at preventing level-1 coercion. Benito illustrated level-3 coercion, and Franklin illustrated level-4 coercion. Obviously, there could be even higher levels. Then Bob suggested the following principle:

Inductive Coercion Principle. It is wrong to perform acts of coercion of an odd number of levels, but it is not wrong to perform acts of coercion of an even number of levels.

The inductive coercion principle would explain why Adolph and Benito's coercion were wrong, but Winston and Franklin's were not. Bob also noticed that the structure of the inductive coercion principle applied to more than just coercion. So he specified a list of *personal harms*, understood as harms to one's mind and body. The list included being killed, disabled, or mentally or physically restrained.[3] He added to the list *harms to property* and he defined *basic harms* to include both personal harms and harms to property. Then he proposed the following principle:

Inductive Harm Principle. It is wrong to intentionally or negligently cause a basic harm (or the risk of a basic harm) of an odd number of levels, but it is not wrong to threaten or to intentionally or negligently cause a basic harm (or the risk of a basic harm) of an even number of levels.

Bob almost immediately realized that the inductive harm principle was too simple. He saw that there was a more complicated idea involved, the idea of a moral right. Attempting to articulate that idea led him to propose a theory of natural rights with three elements:

I. Primary Natural Rights Principle. Everyone has a natural right that others not intentionally or negligently cause them any basic harm (or the risk of a basic harm) and that others not threaten them with a basic harm (or the risk of a basic harm).[4]

The concept of right already contained the inductive structure of the inductive harm principle, because a right entails enforceability. Just to make

the inductive structure clear, Bob made explicit an enforcement exception to natural rights.

II. Secondary Enforcement Provision. When the relevant authorizing conditions are satisfied, everyone has a right to intentionally cause another person a basic harm (or the risk of a basic harm) or to threaten a basic harm as part of proportionate enforcement of a natural right—that is, in order to deter or prevent the violation of a natural right or in order to exact appropriate compensation or proportionate punishment for the violation of a natural right.[5]

In thinking about the state of nature, it seemed obvious to Bob that people could voluntarily waive or transfer their rights by actual consent, so he made explicit one more exception to natural rights, an actual consent exception.

III. Actual Consent Exception. Any person may voluntarily waive or transfer a natural right, either conditionally or unconditionally.

This exception covers a variety of cases, as when a person consents to surgical treatment, or makes a promise, or enters into a mutual agreement.

Libertarianism and the Process of Moral Development

My story of Moses, Fred, and Bob is an oversimplified reconstruction of the development of libertarianism in moral philosophy. The example of the development of libertarianism seems to me to serve as a microcosm of progress in all areas of philosophy. Rather than pause over the details of libertarianism, I want to focus on the process by which it developed. As I reconstruct it, the development of a libertarian theory of natural rights can be understood as part of a larger process of using judgments about particular actual and hypothetical cases to improve our ground-level norms and principles. Call this the *process of moral development*.

The process of moral development involves consideration of actual and hypothetical examples, formulating ground-level norms or principles to explain the examples, and then trying to think of counterexamples to the previously formulated norms or principles as guides to help us formulate more comprehensive ground-level norms or principles. The process involves *equilibrium reasoning*: reasoning that is largely, but not entirely, bottom-up, from judgments about examples to ground-level norms or principles that explain them. In the previous volume, I suggested that this process should be thought of as a process of moral discovery. I contrasted the model of equilibrium reasoning included in the Moral Discovery Paradigm with the model of top-down reasoning that is part of the *Proof Paradigm* (Talbott 2005, 23–35).

As I discussed in the previous volume, the Proof Paradigm seems to me to be a hopeless model for understanding moral development, for it almost inevitably leads to moral skepticism or moral nihilism. One of its collateral effects has been to make it difficult for philosophers to recognize or explain

the process of moral development. The reason is simple. If moral principles were self-evident or provable from self-evident premises, moral development of the kind I have described could not occur. As I have described it, moral development is primarily a process of discovering new actual and hypothetical cases that are counterexamples and thus exceptions to previously accepted norms or principles, followed by attempts to formulate new norms or principles that will cover the new cases. But if moral principles were self-evident or provable from self-evident premises, there would be no counterexamples or exceptions to them, and thus no process of improving them.

There are two ways of understanding the process of moral development, corresponding to two different ways of interpreting the ground-level moral norms or principles that emerge from it. For ease of exposition, I limit myself to principles, with the understanding that parallel distinctions apply to norms:

Categorical Principles. First, the process can be understood as an attempt to formulate categorical principles governed by classical logic. Understood this way, even a single counterexample to an accepted ground-level principle invalidates the principle and requires us to replace it. For example, recognizing that the example of Winston, Adolph, and Anne was a counterexample to the simple prohibition of coercion would make it necessary to discard the simple prohibition. I refer to those who believe that the conjunction of the Primary Natural Rights Principle, the Secondary Enforcement Provision, and the Actual Consent Exception is a categorical ground-level moral principle as *strict libertarians*.

Noncategorical or Defeasible Principles. Second, the process can be understood as an attempt to formulate noncategorical ground-level principles—that is, as ground-level principles that admit of exceptions. When ground-level moral principles are understood in this way, they are taken to hold not categorically, but only *other things being equal*. It is part of the understanding of the principle that there will be exceptions when other things are not equal. If the exceptions are understood noncategorically also, then there is the possibility of a potentially infinite series that starts with a noncategorical principle, followed by a noncategorical combination of the initial principle plus an exception clause, followed by a noncategorical combination of the initial principle with the original exception clause plus an exception clause to the original exception clause, and so forth. When moral principles have this structure I say that the principles are *defeasible*. No matter how many qualifications are built into defeasible principles, they always are understood to allow for more.

Note that if all or most substantive ground-level moral principles are defeasible in this sense, then it is a mistake to regard them as approximations of categorical principles, because there is no finite length categorical principle for them to approximate. Any finite length principle will be understood as holding only other things being equal, because it will be understood to admit of exceptions.

The example of Franklin, Benito, Winston, Adolph, and Anne is the kind of example that leads some people to the conclusion that ground-level moral principles are defeasible. Understood this way, the simple prohibition of coercion need not be discarded. It is assumed to hold only other things being equal.

Ultimately I am going to argue that almost all ground-level moral (and legal) principles are defeasible. Though there may well be some categorical ground-level principles (e.g., It is always wrong to torture young children to death merely for the fun of watching them suffer), I believe that most substantive ground-level moral principles are defeasible. As I explain in chapter 5, there is a categorical meta-level moral principle that explains their defeasibility.

Understood as defeasible principles, the libertarian principles did a better job of articulating the exceptions to the simple prohibition than the coercion minimization principle did. However, it would be expected that there would be exceptions to the libertarian principles, exceptions to the exceptions, and so forth. Understanding the process as one of moral development helps us to understand the attitude that it is appropriate to take toward the ground-level principles that emerge from the process. The only reasonable attitude to take toward them is a nondogmatic one, which regards the currently accepted principles as subject to revision or subject to exceptions. This kind of nondogmatic attitude is an essential part of philosophy that is epistemically modest, that is, that acknowledges the possibility of error and the potential for further improvement (Talbott 2005, 15).

Nozick's Libertarianism

One of the most philosophically sophisticated libertarian theorists was the early Robert Nozick of *Anarchy, State, and Utopia* (1974). Nozick did not claim self-evidence for his libertarian principles. On the contrary, he explicitly constructed his principles to explain particular moral judgments in a variety of actual and hypothetical cases. Nozick's own consideration of examples was much more subtle and nuanced than the simple reconstruction I provided above. I can ignore most of the subtleties and nuances, because I want to focus on the problems that remain. Nozick was aware of the kinds of problems that would be raised for his theory, and he was quite creative in providing solutions to some of them. But even when he attempted to solve them, the structure of his theory prevented him from formulating morally adequate solutions. And in some cases, the structure of his theory prevented him from providing any solution at all. I begin with the former cases.

At the time Nozick wrote his book, there was a well-known example that seemed to raise a serious problem for a strict libertarian view:

Cornelius and the Only Oasis in the Desert, with Slavery Contracts (Nozick 1974, 180). Suppose Cornelius is the owner of an oasis in the desert. Through no fault of anyone, all the other oases in the desert dry up except Cornelius's, which has plenty of water for everyone. On a strict libertarian view, if Cornelius had acquired the oasis without violating anyone's rights (e.g., he did not forcibly appropriate it from someone who properly owned it), it would be morally permissible for him to use whatever force was necessary to prevent people who were dying of thirst from taking his water without his permission and he would be permitted to make his permission conditional on the most extreme terms. For example, he would be permitted to require that thirsty supplicants sign a contract of perpetual slavery in exchange for water. Those who refused to sign would die of thirst.

Most people believe that it would be morally wrong for Cornelius to insist on such onerous terms and that those who voluntarily signed a slavery contract to avoid dying of thirst would not be morally bound by the contract. A theory based on moral judgments about particular cases needs to explain these particular moral judgments. Nozick could have just dogmatically insisted that Cornelius had a right to set any terms he wanted, but that is not the course that he followed. Instead, Nozick incorporated an exception into his account of property rights, the *Lockean proviso*.

The Lockean proviso introduced a minimal consequentialist element into Nozick's theory. It was a proviso on the acquisition of private property that required that private ownership not so disadvantage others that they would be worse off than they would have been in a world with no private ownership at all (*Nozick's baseline*). Nozick believed that his baseline (the level of existence in a world with no private ownership at all) was very low, so that the Lockean proviso would come into play only in cases of catastrophe and the like (1974, 181). Let us agree with Nozick that without private ownership, society would not have progressed above a subsistence level. Then the Lockean proviso would come into play only when private ownership caused someone to drop below that level.

Nozick is correct that adding the Lockean proviso as an exception to strict libertarianism enables him to explain the judgment that it is wrong for Cornelius to set such harsh terms in the example of the only oasis in the desert. Because the Lockean proviso limits Cornelius's property rights to make it impermissible for him to trade water for perpetual slavery and makes invalid Cornelius's contracts of perpetual slavery, even when actually consented to, in Nozick's theory, the Lockean proviso operates as an exception clause to property rights and to the actual consent exception in libertarian theory. Cornelius still owns the water, but he is not permitted to exact such onerous terms in selling it.

Although the Lockean proviso yields the right result in this case, it is difficult to believe that it is the best explanation of the particular moral judgments even in this case. The reason is that Nozick's explanation implies that it would not have been wrong for Cornelius to offer slightly less onerous terms:

Cornelius and the Only Oasis in the Desert, with Perpetual Subsistence Wage Contracts. Suppose that Cornelius does not insist on a slavery contract in return for some of his water, only on a contract to work for him for the rest of one's life at subsistence wages (or at Nozick's baseline, whatever it is). Suppose that before the water shortage, those dying of thirst had all worked at jobs that paid much more than subsistence, so that Cornelius's offer represents a great decrease in their previous quality of life.

On Nozick's libertarianism, even with the Lockean proviso, there would be nothing wrong with Cornelius's offer of a contract of perpetual subsistence labor and he would be justified in using whatever force was necessary to protect his ownership rights to the water from those who were dying of thirst but were not willing to sign his perpetual subsistence level employment contract. Most people's considered moral judgments on this example would conflict with the conclusion of Nozick's theory on this case. It seems that Nozick's own exception to strict libertarianism (the Lockean proviso) is not adequate. But there is more.

Nozick's Lockean proviso comes into play only when it is private ownership itself that is responsible for lowering people below the baseline. When it is due to their own choices, the Lockean proviso does not come into play (1974, 180). Consider another example:

Cornelius and the Only Oasis in the Desert with Slavery Contracts for Those Who Refused to Buy Drought Insurance. If Cornelius had previously offered to sell drought insurance to the people whose oases had dried up—for example, if he had offered to sell them the right to buy water from him at a less exorbitant rate if their oases dried up—and they had refused his offer of drought insurance, then the Lockean proviso would not come into play at all and it would be permissible for him to insist on slavery contracts from those to whom he provided water, because their lack of water would be due not to private ownership, but to their own failure to buy insurance.

Or consider another example that Nozick discusses:

Enclosure Example, with Starvation. While you are asleep one night, your enemy buys all the property around your house and refuses to let you pass over his property. You are trapped in your house. Is it permissible for your enemy to starve you to death by preventing you from crossing his property to get out and preventing anyone who wants to help you from crossing his property to get in? Again, rather than dogmatically sticking to strict libertarian principles, Nozick took such examples to be a test of the adequacy of libertarian theory. He insisted that the "adequacy of libertarian theory cannot depend upon technological devices being available, such as helicopters able to lift straight up above the height of private airspace . . ." (1974, 55, footnote). In this example also, he invoked the Lockean proviso. The Lockean proviso explains why your enemy is not permitted to starve you to death in this way, because it would be a case in which private ownership drove you below Nozick's baseline.

Consider the following variation:

Enclosure Example, with Subsistence. What if your enemy provides you with just enough of the necessities of life to keep you at subsistence level (or just above Nozick's baseline)? Your life would actually be worse than the life of a criminal sentenced to life in prison, because even criminals in prison live lives considerably above Nozick's baseline. However, the Lockean proviso would not come into play, and Nozick's theory would imply it was permissible for your enemy to use whatever force was necessary to prevent you from trespassing on his property. But for most people, the enemy who keeps you trapped living at subsistence level for the rest of your life does you almost as much wrong as the enemy who starves you to death.

Nozick introduces the Lockean proviso into his theory so that it can deal with these exceptions to the actual consent exception. But there was one exception that could not be explained by the Lockean proviso or any other part of his theory that he nonetheless allowed. This exception covers natural rights infringements in cases in which it is not possible or is too costly to obtain a property owner's consent in advance, but in which compensation can be paid after the fact.

Absent Neighbor. Consider, for example, a situation in which the only way to save your child's life is to rush him to the hospital, and the only car available is your neighbor's. Your neighbor is currently away from home, but fortunately he left the key in the ignition.[6] Even without your neighbor's permission, Nozick believes that you are permitted to use his car and negotiate compensation later. Here is what Nozick says about such examples:

> The reason one sometimes would wish to allow boundary crossings with compensation (when prior identification of the victim or communication with him is *impossible*) is presumably the great benefits of the act; it is worthwhile, ought to be done, and can pay its way. But such reasons sometimes will hold, as well, where prior identification and communication, though possible, are more costly even than the great benefits of the act. Prohibiting such unconsented-to acts would entail forgoing their benefits, as in the cases where negotiation is impossible. (1974, 72–73)

Nozick never does find a principle to cover this kind of exception. Even his description of the situations invites a consequentialist explanation. However, these kinds of situations actually raise a deeper problem for Nozick's theory. To see why, consider the following variation on the example:

Hard-Bargaining Neighbor. Suppose that your neighbor had been standing next to his car, so that compensation could be negotiated. Suppose that your neighbor has no plans to use his car that day, but seeing that you child's life is in danger, he refuses to consent to your using his car for anything

less than $1,000,000. Suppose, also, that your neighbor knows that, although your assets total less than $1,000,000, you would be willing to commit to pay $1,000,000 in installments if necessary, to save the life of your child and there is no other way to get your child to the hospital. On Nozick's account, when negotiation is not impossible and not costly, you would not be permitted to take your neighbor's car without agreeing to pay him $1,000,000. This would not fit with most people's particular moral judgments. In such a circumstance, most people would not think that you needed your neighbor's consent to use his car. If you could get away with it, you would be permitted to take your neighbor's car without his consent and then to return it later and to pay your neighbor a reasonable amount for his loss, not the $1,000,000 that it would have been necessary to pay to obtain his consent.

Nozick's own discussion of the cases in which prior negotiation is impossible or too costly opens the door to a consequentialist or hypothetical consent exception to the actual consent exception. Once the door is open, it becomes apparent that there are many more cases that would fall under it. To see why, consider an elaboration of an example introduced by Nozick himself to explain why the use of a baseline in the Lockean proviso did not make his theory consequentialist:

Medical Researcher. Marie is a researcher who invents a cure for an otherwise fatal illness from easily available materials (e.g., from water and carbon dioxide). Marie is the only person who knows how to make the cure, but everyone has access to the materials from which it is made. A pandemic of the fatal illness spreads throughout the earth. Marie agrees to supply the medicine only to those who agree to be her slaves for life. Those who do not agree to her terms die. So before long, everyone else on earth is contractually bound to be Marie's slave for life.[7]

On Nozick's theory, Marie has the moral right to her slaves' services and none of her slaves has a right to rebel. Examples like this illustrate the fact that Nozick's theory is a historical theory. Moral permissibility depends on the actual history. Even a slave society can be justified, given the appropriate history, as in this example. Nozick used this sort of example to show that his theory was not consequentialist. However, the argument cuts both ways. For most people, the example shows a flaw in Nozick's own theory. Though Nozick did not seem to be aware of it, through the kind of bottom-up reasoning that he himself employed, the example of the medical researcher should have led him to doubt his theory.[8]

Voluntariness

One way of trying to resolve the example of the medical researcher is to deny that the consent of the slaves is given voluntarily. The idea is that when one faces a choice between death and perpetual slavery, there is no real choice, so the consent is not voluntary. This is not a move that would be available to

Nozick, because he explicitly disavows a conception of voluntariness that depends on the number of available choices (1974, 262–265). In any case, it is not a very promising direction in which to look for a solution. The reason is simple. Marie the researcher is not responsible for reducing the number of choices available to others; she has *increased* them. Before her discovery, there was no alternative to death for those who contracted the disease. After her discovery, they have a new alternative, perpetual slavery. For any theory based on autonomy, it would seem that Marie has *enhanced* the autonomy of those who contract the disease. At least, she has in no way diminished it. And so, Nozick's theory implies that Marie does nothing wrong, not even if she enslaves every human being on earth.

Nozick at least tried to solve the puzzles generated by the example of the only oasis in the desert and by the enclosure example. Nozick had no solution to the puzzle generated by the example of the medical researcher, so he simply insisted that there is nothing wrong with what she does. All of these examples point to the need for further exceptions to Nozick's libertarian rights, and at least strongly suggest that their rationale will be consequentialist.

The Evaluation of Particular Cases and the Phenomenon of Theoretical Inertia

It might seem that these sorts of examples are too far-fetched to be taken seriously. That was not Nozick's attitude. He took such examples very seriously and expected his opponents to do so too. For the purposes of philosophical understanding, hypothetical examples often have an advantage over real-world examples, because they can be fashioned to focus on important elements of the theory to be evaluated that might otherwise be obscured by the details of real-world examples. In any case, as Nozick realized, there were real-world examples that closely resembled some of the hypothetical ones.

Consider the example of the medical researcher. It is easy to find real-world analogues. Pharmaceutical companies selling lifesaving drugs have been willing to charge prices for the drugs that will bankrupt an uninsured patient. The results are that pharmaceutical companies are some of the most profitable companies. According to Nozick's theory, they have every right to those profits, even if they bankrupt those who need the drugs, because, according to Nozick's theory, not even perpetual slavery would be too high a price. Should Nozick have given up his libertarian principles in the light of such examples?

How do we tell when a particular moral judgment should be given more weight than a moral principle with which it conflicts? There is no decision procedure for such a determination. It involves an exercise of moral judgment. In this book, rather than try to describe how to do it, I try to provide lots of examples and rely on your ability to do it. If we were not able to do it in at least some cases, moral inquiry would quickly come to a halt.

Even though there is no decision procedure for establishing equilibrium among our moral principles and our particular moral judgments, there is one phenomenon that we need to be aware of, the *phenomenon of theoretical inertia*. This is the phenomenon that the very adoption of a theory makes one less able to appreciate the force of particular cases that potentially conflict with it. A corollary of this phenomenon is that the advocates of a philosophical theory almost always accord less significance to potential counterexamples than they should, and the advocates of opposing theories more. Therefore, in assessing the weight to be given to judgments about particular cases, the judgments of knowledgeable observers with no stake in the competing explanatory theories are generally a more reliable guide than the judgments of those with a stake in one of the competing explanatory theories. I discuss how the phenomenon of theoretical inertia applies to my own account in chapter 14.

The phenomenon of theoretical inertia seems to me the best explanation of the early Nozick's willingness to allow his libertarian theory to overrule most people's particular moral judgments in the examples discussed above. Those examples were not marginal or controversial examples. They are central examples from the history of moral development. In Hugo's *Les Miserables*, the pivotal event is Jean Valjean's theft of bread to feed his starving children. It would not have made the story any less forceful if Valjean had stolen medicine to keep his children from dying.

When Lawrence Kohlberg (1981) constructed his theory of moral development, one of the tests he employed to determine the level of moral development was the *example of Heinz and the druggist*. Heinz needs a drug to save his wife's life, but he doesn't have enough money to pay for it. The druggist refuses to part with the drug on any other terms than a full-price sale. But Heinz can steal it. Kohlberg ranks a subject's moral development in part based on the ability to explain why Heinz should steal the drug. Of course, the case for stealing the drug would be even stronger if the druggist refused to sell the drug to Heinz unless Heinz agreed to be his slave for life. And when Gilligan (1982) proposed an alternative to Kohlberg's theory of moral development according to which Heinz should explore other alternatives to stealing the drug, she never would have suggested that one of the alternatives he should consider was letting his wife die or agreeing to be the druggist's slave for life.

If theoretical inertia prevented the early Nozick from appreciating the force of these particular moral judgments, it is to his credit that he was later able to overcome that inertia. Later in his life, Nozick augmented his libertarian principles to allow for laws that would prohibit the kinds of wrongs illustrated by these examples. His explanation of the change was in terms of the symbolic value of such laws (1989, 291–292). I believe there is a need for a deeper explanation. In any case, Nozick provides us with an example of what political philosophy is like when it is done nondogmatically. In spite of the phenomenon of theoretical inertia, Nozick was able to feel the force of particular

moral judgments that conflicted with his theory and changed his theory to accommodate them.

Beyond Libertarian Accounts

As I reconstruct it, libertarianism has three parts: a primary natural rights principle, a secondary enforcement provision, and an actual consent exception. Strict libertarians interpret the conjunction of the three principles categorically, as admitting of no exceptions. Nozick departed from strict libertarianism by adding the Lockean proviso to address the example of the only oasis in the desert and the enclosure example and by allowing rights infringements without consent when negotiating consent would be impossible or too costly. I have suggested that Nozick's exceptions are not adequate. At this point, there are two ways to proceed:

(1) *Ground-level explanation.* We can continue to articulate exceptions to the actual consent exception and then go on to consider exceptions to the exceptions and exceptions to the exceptions to the exceptions, and so forth. This is not the path I will follow, because I believe that there is no finite end to the process of adding exceptions, and exceptions to the exceptions, and so forth. Why not? Why, as so many philosophers have pointed out (e.g., Ross 1930; Dancy 2004), is it so hard to find interesting, substantive, ground-level norms or principles that are categorical (i.e., exceptionless)?[9] Why are most, if not all, interesting, substantive, ground-level moral principles defeasible? I answer these questions in chapter 5. To do so, I will have to move up a level.

(2) *Meta-level explanation.* At this level, we are not trying to articulate principles of ground-level moral reasoning; we are trying instead to explain the moral appropriateness of ground-level moral principles and of potential exceptions to them. At this level, we attempt to articulate a meta-level principle that will explain not only the actual consent exception, but also the exceptions to the exception, the exceptions to the exceptions to the exceptions, and so forth. This is the path that I find most promising.

What about the meta-level principles? Are they to be understood to be defeasible, also? Should we expect to find a potential infinity of morally appropriate exceptions and exceptions to exceptions, and so forth at the meta-level as well as at the ground level? And if so, won't there be a need for third-level (meta-meta-level) principles to explain the morally appropriate exceptions to the meta-level principles and then a fourth level and a fifth level, and so again to infinity?

Perhaps surprisingly, I believe that the answer to all these questions is no, because at the second level we find an exceptionless meta-level principle that explains the potential infinity of morally appropriate exceptions at the first level. Because we have no infallible insight into meta-level moral principles, our attempts to formulate the relevant meta-level explanatory

principle depend on our being able to recognize enough exceptions to ground-level moral principles to support an explanatory theory at the meta-level. Whatever principles we do formulate at the meta-level will be fallible and subject to correction, but that does not imply that they will be defeasible. Consider an analogy. It is possible that there are laws of physics that hold everywhere in the universe. If so, the fact that our beliefs about the identity of those laws of physics are fallible is compatible with the laws themselves being exceptionless.

Contractarianism as a Nonconsequentialist Meta-Theory

How are we to explain the exceptions to Nozick's ground-level libertarian principles? A natural suggestion is that the exceptions to Nozick's theory discussed previously can all be explained by a single nonconsequentialist *contractarian* principle. *Contractarians* propose to explain all ground-level moral principles in terms of some sort of *hypothetical consent*, for example, Rawls's account based on consent in the *original position* (1971, chap. 3; 1993, 22–28) or Habermas's account of consent in an ideal rational discourse (1990, 58).

Consider, for example, how a contractarian account that employs an original position behind a veil of ignorance could address the example of the medical researcher.[10] Behind a veil of ignorance, I would not know whether I was the medical researcher who had developed the cure or one of the potential purchasers of the cure who would die without it. Behind the veil, it seems that everyone would agree to limits on how much the medical researcher could charge for the cure, and would never allow the researcher to require perpetual slavery contracts.

It is easy to see how a hypothetical consent account could also provide a satisfactory explanation of exceptions to cover the other examples, the Cornelius examples, the enclosure examples, and the examples of the absent neighbor and the hard-bargaining neighbor. So hypothetical consent theories seem to be strong candidates for the meta-theory we are looking for. However, there is a puzzle about all such hypothetical consent accounts, a puzzle that ultimately leads beyond them, also. The puzzle is this: Actual consent is something that actually takes place. To determine the terms of an actual agreement, one must conduct a historical investigation. But how does one determine the terms of a hypothetical agreement?

This would not be a problem if by hypothetical agreement we simply meant *would agree if asked*—for example, if the claim were the factual claim that, if she were asked, the medical research Marie would agree that you are permitted to steal her cure. But, of course, this is not the relevant kind of hypothetical consent, because it would not explain why it is permissible for you to steal Marie's cure, even in a case in which she would not consent to the theft if she were asked to do so. The exception covers that case, too.

Of course, for both Rawls and Habermas, the relevant kind of hypothetical consent is much more abstract. Marie may admit that behind the veil of ignorance she would agree that there should be limits on what people like her can charge for lifesaving medicines, but insist that that makes no difference to what she should agree to in her current situation, in which she knows who she is.

Historically, an actual consent exception to primary ground-level moral norms and principles arose long before hypothetical consent exceptions were contemplated. This made it seem as though the actual consent exception was fundamental and that hypothetical consent somehow inherited its moral significance by being a kind of consent. However, there is an insuperable problem with this way of understanding the moral significance of hypothetical consent. Hypothetical consent is not a pale, less binding form of consent. It is no consent at all (R. Dworkin 1977, 151).

Although it is almost irresistible to think that the moral force of hypothetical consent is somehow derived from the moral force of actual consent, the opposite is closer to the truth. It is more accurate to say that the moral significance of actual consent is derivative from the fact that an actual consent exception to libertarian natural rights would be agreed to under the appropriate circumstances (e.g., the original position or the ideal speech situation). But even this formulation invites misunderstanding, because it makes it seem that the fundamental explanatory level is the level of explanation in terms of hypothetical consent. This is a mistake.

To see that there is a more fundamental level of explanation, consider how advocates of hypothetical consent theories typically determine what people would consent to in the relevant circumstances. One way to try to answer this question would be to make a very thorough study of human psychology on the basis of which to be able to predict what individuals would or would not agree to under a variety of hypothetical situations. This seems a hopeless project, because there is so much psychological diversity among individuals that for all but the most clear-cut decisions, it would be nearly impossible to predict what any representatively diverse group of people would agree to.

In any case, hypothetical consent theorists never undertake this kind of detailed psychological study. Instead, they typically follow Rawls's model and try to determine the terms of the relevant hypothetical contract by asking what terms it would be rational (or reasonable) to agree to. However, if some agreements can be determined to be reasonable and others unreasonable simply by thinking about the reasons for and against them, the obvious question is: What makes the reasonable ones reasonable and the unreasonable ones unreasonable (Thomson 1990, 360)?

Although it is theoretically possible that the answer to that question be given in terms of hypothetical consent (a higher level theory of what terms are reasonable in lower level agreements), that would seem to simply push the problem up a level. We would then have to wonder how we could ever know which potential upper-level agreements would be reasonable and which

unreasonable. At some level, there would have to be some way of distinguishing between reasonable and unreasonable agreements that was itself not based on hypothetical consent. That would be the fundamental level of explanation. Hypothetical consent would then turn out to simply be a useful heuristic for detecting what really made agreements reasonable or unreasonable. This is, in fact, an important role that hypothetical consent does play in my consequentialist theory: Hypothetical consent in an original-position-type situation is a useful heuristic for gauging the equitable promotion of well-being.

Habermas tries to avoid this objection by refusing to predict the results of the hypothetical process, the process of ideal rational discourse, in advance. He insists that the processes of argumentation and negotiation among individual parties whose interests conflict "must actually be carried out" (1993, 16). But, of course, because ideal rational discourse is an *idealization*, it *can't* be carried out. Any real-world process will have to be evaluated by the extent to which it approximates the ideal. But how could the comparison be made, if the idealized process can never be carried out? All we can ever determine are the results of real-world processes. To project the results of the ideal process on the basis of a real-world process, we would have to be able to use the real-world process to predict which kinds of considerations would be decisive in the ideal process. This is just another way of saying that we would have to be able to project what it would be reasonable to agree to in the ideal process, on the basis of what it is reasonable to agree to in the real-world process. We could never do this if we didn't have some way, fallibly of course, of being able to recognize when the results of real-world processes were reasonable that did not require comparison with the ideal process.[11]

What about the real-world process? Is it always necessary to run the real-world process to determine what it would be reasonable to agree to in that process? There is something morally attractive about this position, because it would provide an antidote to the almost irresistible presumption in philosophy that all reasonable people would agree with me. It really is important that everyone affected be recognized as having a contribution to make in determining what is reasonable.

Nonetheless, it is too extreme to hold that we can never recognize terms of reasonable disagreement on moral norms without engaging all those affected in a real-world process of more or less rational discourse. Habermas himself implicitly acknowledges this when he offers human rights norms as an example of the kinds of norms that would be agreed to in ideal rational discourse (1990, 105). How could he think that the real-world process of discourse that has generated human rights norms approximated the ideal process when the great majority of the world's population has never contributed to the discourse on human rights? If such a discourse did take place, we know that many people would object to women's rights on religious grounds. How can he predict the results of this discourse? The only way I can answer that question is to think that he can himself judge which reasons are weightier. He does

not need to carry out either the ideal process or the real-world process to make judgments about the weight of reasons, at least in some cases.

I have to admit that if Rawls or Habermas or someone else defined a process that yielded conclusions about justice that made it independent of the equitable distribution of well-being, and if people generally agreed that the results of the process were just, then I would have to reconsider my consequentialism. That has not happened. When metaphysical Rawls formulated his version of the original position thought experiment, he argued that it would lead to consensus on the consequentialist maximin expectation principle as a general conception of distributive justice (1971, chap. 3).

The consequentialist element is even more explicit in Habermas's account, because for Habermas, the moral question to be decided through rational discourse is this: What norms would be "equally good for all" (1993, 59)? This led Lafont to argue that Habermas is a substantive rather than a procedural realist about justice, because his own view is that justice is what is equally in everyone's interest (1998, 68). Habermas replied that what is equally in everyone's interest is not a fact; it is the result of a process of deliberation (2003, 266–267). Why is he so sure that there is no formula for what is "equally good for all" or for equitably promoting well-being that would explain what the parties engaged in deliberations about justice are trying to figure out? Whether or not there is such a formula seems like the kind of issue that it would be best to leave to the real-world process of rational discourse itself rather than to try to settle in advance.

So the success of hypothetical consent theories cries out for an explanation at a deeper level. The deeper level would explain why using some sort of ideal impartial (Rawls) or intersubjective (Habermas) hypothetical agreement would be a good test for determining the principles of morality or justice. I believe there is a deeper level of analysis and that, at that level, the explanation is consequentialist. At this deeper level of analysis, there is a consequentialist principle (the main principle) that explains the moral appropriateness of exceptions to our ground-level moral norms. It will explain the moral appropriateness of exceptions to the libertarian rights, including the actual consent exception itself. It will also explain the usefulness of idealized impartial or intersubjective consent tests for determining which exceptions are appropriate (including the actual consent exception itself). So the order of explanation is that the consequentialist main principle explains the usefulness of a hypothetical consent test for moral appropriateness and then that hypothetical consent test explains the moral appropriateness of exceptions to libertarian natural rights, including the actual consent exception itself.

As I explain in chapter 9, the moral significance of consent—both hypothetical (as a test of the moral appropriateness of exceptions to ground-level moral norms and principles) and actual (i.e., the moral appropriateness of the actual consent exception to libertarian natural rights)—is due to the fact that, given certain background assumptions, consent is a reliable indicator of improvements

in well-being. Or, to put it another way, if consent were not a reliable indicator of improvements in well-being, it is unlikely that there ever would have developed an actual consent exception to libertarian natural rights or hypothetical consent tests for ground-level norms of morality and justice.

The Actual Consent Exception to Libertarian Natural Rights Is Too Broad and Too Narrow

Understood categorically (i.e., as not admitting of exceptions), the actual consent exception to libertarian natural rights is both too broad and too narrow. We have already seen how it is too narrow, because we have seen lots of examples in which it would be permissible to infringe others' rights without obtaining their consent. We can use slight variations on these examples to show that the actual consent exception is also too broad. The cases in which it is too broad also illustrate an important parallel between legal principles and moral principles.

Consider the examples involving contracts of slavery or perpetual subsistence earnings discussed above. If you had entered into any of those contracts in the United States, you would be able to go to court to have them voided as *unconscionable*. Unconscionability is a legal doctrine that developed as an exception to the rule that contracts are to be enforced as written.[12] The doctrine did not originate in legislation. It was a doctrine introduced by judges to justify not enforcing contracts as written in certain extreme cases. The unconscionability doctrine is an exception to the actual consent exception to libertarian natural rights. Even with the addition of the Lockean proviso, we have seen that there are still exceptions to the actual consent exception (e.g., the example of the medical researcher). These examples show that the actual consent exception to libertarian natural rights is too broad, even when qualified by the Lockean proviso, because these are cases in which people would not be morally (or legally) bound by their actual consent.

How should we understand the unconscionability doctrine? Is it to be understood as a nonconsequentialist doctrine that rules out the agreements on the grounds that they are not truly voluntary? Is it to be understood as a consequentialist doctrine aimed at avoiding really bad results? It is too early for me to try to answer that question. I return to it in chapter 9.

Although the unconscionability doctrine is a legal doctrine employed by judges, it has a precise parallel in moral thought. Even in a state of nature with no legal system to enforce contracts, it is plausible to think that consent would be morally binding. Even so, if in the state of nature you had consented, by giving your word, to be Marie's slave for life in return for a dose of her lifesaving medicine, you would not be morally bound by your agreement. There would be no moral requirement for you to keep your word. If you had an opportunity to escape or to participate in a slave revolt, there would no moral requirement not to do so. So there is an unconscionability exception to the actual consent exception in both morality and law.

The examples discussed above illustrate another legal doctrine. If we focus not on the contracts, but on the acts that would be necessary to avoid having to enter into an unconscionable contract, we find a moral parallel to the necessity defense in the law. For example, if stealing food were necessary to prevent your children from starving, you would have a legal defense to criminal prosecution for stealing, which in Anglo-American law is called the *necessity defense*. The general formula for the necessity defense is that it is a defense that applies when the harm avoided by breaking the law is greater than the harm caused by breaking it.[13] The necessity defense could be invoked in any of the previous examples to justify stealing rather than death or perpetual slavery or perpetual subsistence. The necessity defense applies more broadly than to cases in which one of the alternatives is an unconscionable contract. For example, it applies to the example of the absent neighbor, in which it is necessary for you to take your neighbor's car without consent, not because the proposed terms of consent are unconscionable, but because your neighbor is not available to give consent.

Let us call the cases to which the necessity defense applies *necessity exceptions*. So now we have discovered two categories of exception to both moral and legal norms, necessity exceptions and unconscionability exceptions. Necessity exceptions show how the actual consent exception to libertarian natural rights is too narrow; unconscionability exceptions show how it is too broad. However, we have not come close to exhausting the variety of ways that it is too broad or too narrow. It turns out that the most important category of exceptions to libertarian natural rights are examples that show that the actual consent exception is too narrow. This is the category of legal solutions to collective action problems. When a morally appropriate system for enacting legal solutions to collective action problems enacts such a law, the law is typically binding on all citizens, whether or not they explicitly consented to the law and whether or not they explicitly consented to the legal system that produced the law. Actual consent is not necessary. Hypothetical consent, of the relevant kind, is enough. And there is a consequentialist explanation of why. I return to this topic in the next chapter.

A Parallel between Morality and the Law

The example of the unconscionability exception in both morality and law illustrates a strong parallel between the two domains. This parallel reveals a deep unity between the situation of an individual deciding whether to make an exception to a moral principle and the situation of a judge deciding whether to make an exception to a legal principle. As we will see in chapter 5, laws and legal principles have the same kind of defeasibility as ground-level moral principles. The very same consequentialist meta-principle explains why ground-level moral and legal principles are both defeasible, and determines when making an exception to them is morally appropriate.

Although ultimately, as I explain in chapter 5, there is a deep unity that explains the parallel between moral systems and legal systems, it is often useful to step back and view them as separate systems. When we do so, we often find useful parallels, illustrated by the parallel between necessity and unconscionability exceptions in morality and the law.

Conclusion

In this chapter I have used a dialogue to recapitulate the history of moral development that led to the development of libertarian natural rights theory, as represented by Locke and by Nozick, not because libertarianism is a universal stage of moral development, but because I want to show how the process that leads to libertarianism leads beyond it. Even if there were a time at which it was morally appropriate to think of ourselves as having libertarian natural rights, the main principle explains why it was an improvement to replace libertarian natural rights with a different and more extensive set of universal human rights.

THREE

The Main Principle

In the previous chapter, I explained how the process of moral reasoning that leads to libertarian natural rights principles also leads beyond them. I continue to use exceptions to libertarian natural rights as an expository device, because, ultimately, it helps me to explain why the human rights that should be universal are different from and, especially, more extensive than libertarian natural rights.

However, there is a danger that this expository device could be misunderstood as limiting my account of human rights only to those who begin with libertarian natural rights in a state of nature. This would be a mistake. The main principle is a meta-theoretic principle that will explain the exceptions to any ground-level moral practice.

The near opposite of a libertarian starting point would be a starting point that begins with a moral code that authorizes a sovereign to exercise absolute coercive power. If we adopt this starting point, we can tell a story about how limits on the sovereign's power can be morally appropriate. This is much closer to the historical story of the development of human rights in Western Europe than a fanciful story that starts in a state of nature. In the previous volume (Talbott 2005, chap. 4), I discussed a striking example of this sort of development, the example of Bartolomé de las Casas. Las Casas originally helped to colonize the Americas and thus to bring them under the legal authority of the king and queen of Spain and the religious authority of the Pope. Ultimately, he decided that the imposition of both kinds of authority on the natives was a disastrous mistake. Las Casas came to believe that the American natives should have been allowed to have their own government and practice their own religion. This change in the moral views he had held when he first arrived in the Americas was endorsed by the main principle.

The very same principle that explains the exceptions to libertarian natural rights will also explain exceptions to moral codes that authorize absolute sovereigns and, indeed, exceptions to all other moral codes in traditions that have passed the consequentialist threshold. It will also explain why, regardless of starting point, moral progress leads toward legal guarantees for a set of universal human rights. I use the example of exceptions to libertarian natural rights to illustrate one pathway to legal guarantees of human rights. It is a useful expository device, because, although there is no longer any serious moral defense of absolute sovereigns, moral defenses of some form of qualified

libertarianism are common and are often used as a basis for opposing human rights, especially economic and social rights.

In this chapter, I continue the bottom-up exposition from the previous chapter. Rather than attempt a full statement of the main principle, I start with the core idea and then gradually construct elements of the principle by considering what kind of principle would explain the moral appropriateness of a variety of exceptions to libertarian natural rights. Not until the end of the chapter will I be able to state a preliminary and then a final version of the main principle.

Because my account is consequentialist, I have to worry that I am going to repeat one or more of the mistakes the utilitarians made. So it useful to begin with a consideration of what we can learn from the failure of utilitarianism.

What Else We Can Learn from the Failure of Utilitarianism

I have already identified the most important lesson to be learned from the failure of utilitarianism: that the distribution of well-being matters. The main principle does not have utilitarianism's distributional blind spot. It favors the *equitable* promotion of well-being. To avoid confusion, it is important to understand the role of equity in the main principle. In ordinary language, equity can be used in two different ways, to refer to fairness in *distribution* or fairness in *procedures*. In the main principle, *equity* is used in the first sense. It is a feature of distributions of well-being.

There are two other lessons to be learned from the failure of utilitarianism. The second lesson is an epistemological lesson. Utilitarianism is an optimizing theory. The epistemological lesson is that human beings will never know what act or system of rules or social practices would be optimal. So if utilitarianism requires us to choose the optimal act or to act in accordance with the optimal set of rules or social practices, we might as well give up. We are sure to fail.

Utilitarians sometimes try to address this problem by distinguishing objective and subjective utilitarianism. Subjective utilitarianism would require only acting in ways that one *believes* to be optimal or acting in accordance with the rules or social practices that one *believes* (or is justified in believing) to be optimal. But this is no help. We *know* that it is extremely unlikely that any human being would ever be lucky enough to hit upon an optimal act or optimal system of rules or social practices, so, unless we are prone to self-deception, we will never *believe* of any act or system of rules or social practices that it is optimal. So we can't even satisfy the subjectivized version of utilitarian theories. Again, we might as well give up. This is a decisive objection to either the objective or subjective version of Brandt's (1992) ideal rule utilitarianism, that would have us act in accordance with the "ideal" system of moral rules, where a system of moral rules is *ideal* "if its currency in [our] society would produce at least as much good per person . . . as the currency of any other moral code" (1992, 119).

The main principle avoids this problem because it is a comparative principle. In a typical case, there will be a moral status quo option and one or a small number of other options for exceptions to the status quo. The main principle favors the option that does the best job of equitably promoting well-being. I express this by saying that the main principle is a *comparative* rather than an *optimizing* principle.

We could make a similar move to make Brandt's theory a comparative one by limiting the alternatives to a given set of relevant alternatives. Then we could define a comparative, subjective version of Brandt's ideal rule theory: Given a set of relevant alternative moral codes, act in accordance with the member *of the given set* that one *believes* to be maximal.

This leads to the third lesson that we can learn from utilitarianism, and especially from Brandt's theory. In Brandt's theory, when we evaluate moral codes, we consider what the effects would be of its "currency in society." This "currency" provision of Brandt's theory generates what I call a *coordination problem*, which is a problem even for the comparative, subjective version of Brandt's theory. Here is an illustration of the problem:

Suppose that you live in a society in which there is a norm to drive on the right. Suppose that you perform a psychological experiment that shows that there would be fewer and less severe auto accidents, and thus that overall utility would be greater, if everyone drove on the left. Then the comparative, subjective version of Brandt's theory would require you to follow the rule of driving on the left, because you believe it would be maximal if it were generally complied with, *even if you know that everyone else is driving on the right*. Clearly, in such a circumstance, it could well be disastrous for you and those around you if you were to unilaterally change your driving practices and start driving on the left. Even a comparative, subjective version of Brandt's theory is inadequate, because it fails to solve this coordination problem.[1]

This coordination problem is a reminder that ground-level moral norms and principles coordinate expectations and behavior and that this kind of coordination is responsible for a good part of their effectiveness in promoting well-being. The third lesson to be learned from the failure of utilitarianism is that any adequate consequentialist meta-principle for explaining the moral appropriateness of changes in ground-level moral thought will have to address this coordination problem. The main principle addresses this issue by evaluating any proposed change in a moral practice in two parts: as a substantive moral practice and as a practice of implementation. To be endorsed by the main principle, the practice of implementation usually must solve a coordination problem.

Explaining the Moral Appropriateness of Exceptions to Libertarian Natural Rights

At its core, the main principle is a general principle for ranking systems of social practices (i.e., complete systems of social practices, including a legal

system, an economic system, an educational system, a system for producing and caring for children, etc.) on the extent to which they equitably promote the well-being of those who are subject to the practices. The core ranking principle is not a principle of personal morality; it simply generates a moral ranking of systems of social practices based on the level and distribution of well-being that they would generate. I say more about the key terms (e.g., *equity* and *well-being*) shortly.

Because all human societies that have ever existed have had systems of social practices that are far from optimal as measured by the main principle's core ranking, the most important role of the core ranking is that it provides the basis for an evaluation of potential improvements to the existing system of practices. In the real world, potential improvements are never optimal. Usually there is a relatively small set of available options. In such a context, the main principle's core ranking principle generates a ranking of the relevant alternatives. It will take this chapter and the next for me to fully explain how it does so. I begin with the key idea and then progressively develop and deepen it.

In the remainder of this chapter, I use exceptions to libertarian natural rights as clues to important features of the main principle. Because the main principle must be able to explain the moral appropriateness of the exceptions that are appropriate (and the inappropriateness of those that are inappropriate), we can use the examples to reason abductively to conclusions about the content of the main principle. Because I will be considering exceptions to libertarian natural rights, it is useful to repeat here the libertarian natural rights principles:

Libertarian Natural Rights (With Enforcement Provision and Actual Consent Exception). Everyone has a natural right that others not intentionally or negligently cause them a basic harm (or the risk of a basic harm) and that others not threaten them with a basic harm (or the risk of a basic harm). When the relevant authorizing conditions are satisfied, everyone has a right to intentionally cause another person a basic harm (or the risk of a basic harm) or to threaten a basic harm as part of proportionate enforcement of a natural right—that is, in order to deter or prevent the violation of a natural right or in order to exact appropriate compensation or proportionate punishment for the violation of a natural right. Finally, any person may voluntarily waive or transfer a natural right, either conditionally or unconditionally.

When we look at the variety of morally appropriate exceptions, exceptions to exceptions, and so forth, to libertarian natural rights, it is hard to believe that most or all of them could be explained by a principle that ranks social practices on the basis of only two factors, the overall amount of well-being and its distribution. So I begin with a series of examples that at least make it plausible that those two factors are decisive in some cases. The examples will also lead us to a fuller articulation of the main principle's standard for moral improvement. As in the previous chapter, the examples compress

hundreds, if not thousands, of years of moral development. I assume that all of the examples take place in a state of nature in which there is general agreement on libertarian natural rights as ground-level moral principles. Thus, there is general agreement that everyone has a right not to be coerced. The examples are cases that raise the possibility of exceptions to that prohibition.

Initial Wild Beast Example. To make the examples simple, I assume that the relevant facts are common knowledge. In the first example, you and I are among a group of four people trapped in a cave by a wild beast. If we do nothing, the beast will eat one of us each day for the next four days. We want to try to escape before the beast gets hungry. Because the cave opening is so narrow, if we try to escape, we will have to escape in single file. The beast is waiting outside. The first escapee through the opening will attract the attention of the beast and be pursued by the beast. The probability of the first escapee's successfully escaping the beast is very low, only .05. However, the remaining three will be able to easily escape while the beast pursues the initial escapee. The probability of each of the remaining three successfully escaping is 1.0. The four of us are deliberating together about what we should do.

Libertarian natural rights would permit one of us to volunteer to go first, thereby risking her life to save the rest. Suppose no one volunteers. Then someone proposes the following:

Equal Chance of Selection Practice. The proposal is that if at least three of us agree, we use a random coin-tossing device (which we happen to have with us) to construct a fair procedure for selecting who will go first from among the four of us.[2] Each of us will have a 1 in 4 (.25) chance of being selected to go first. If we follow this practice, each of us will have a probability of .7625 (= .75(1) + .25(.05)) of surviving. Suppose that the other three of you agree to the equal chance of selection practice, but I do not. Instead, I propose the following alternative:

Unequal Chance of Selection Practice. In this practice, the random coin-tossing device is used to construct a fair procedure for selecting who will go first from among the three of you. Each of you will have a 1 in 3 chance of being selected, thereby guaranteeing that I will successfully escape. In the language of collective action problems, the unequal chance of selection practice would allow me to *free ride* on your cooperation. If you adopt this practice, I will be sure to survive and the rest of you will have a probability of .6833 (= 2/3(1) +1/3(.05)) of surviving, less than the probability of .7625 that each of us would have under the first practice, but still much better than certain death if we all remain in the cave.

This simple example can be used to illustrate how practices are substantively evaluated by the main principle. However, it will also introduce a complication that will require a more complex kind of evaluation, as I explain shortly. First, the substantive evaluation of the two alternatives in comparison

to each other and to the status quo. In a substantive evaluation, changes to the status quo are evaluated on the basis of their effects if they were successfully implemented. Clearly, if implemented, either practice described above will improve the chances of survival of all four of us over the status quo, which is assumed to allow no exceptions to the libertarian prohibition on coercion, because either practice will, at the time that the coin is tossed, give all of us a higher probability of surviving than simply to act on our own. However, it requires no theory of equity to realize that a practice that gives everyone a 76% chance of surviving (the equal chance of selection practice) is more equitable than a practice that gives one person a 100% chance and everyone else a 68% chance of surviving (the unequal chance of selection practice). In the next chapter I introduce an expanded original position test as a heuristic for making comparisons of this kind. In this chapter, because the cases are so clear, I just rely on intuitive judgments.

Even this simple example shows something important about how the main principle evaluates social practices. It evaluates them not on the actual well-being they produce, but on some probabilistic measure of expected well-being. To capture this idea, I will say that the main principle evaluates well-being in terms of *life prospects*. Thus, the main principle substantively evaluates alternative moral practices on the extent to which they equitably promote life prospects. I say more about how life prospects are compared in the next chapter.

In the initial wild beast example, probability of survival serves as a proxy for this measure, whatever it may be. Often I use life expectancy as a proxy for life prospects. Of course, life expectancy would not be a good proxy in all circumstances (yet another generalization with exceptions). However, in some examples, it is an adequate measure.

The initial wild beast example also illustrates the fact that it matters *when* life prospects are evaluated. Suppose that the equal chance of selection practice is adopted. The coin is tossed, and I am selected to exit the cave first. After I have been selected, I could rightly claim that the practice had not improved my life prospects at all, because at that time, as the three of you are starting to push me out of the cave, my chance of survival is no different from what it was before the procedure was implemented. Why should it matter what my life prospects were before the coin was tossed? It seems obvious that it does matter, but it is not easy to say why. I say more about this question in the next chapter.

The Coordination Problem

Finally, the initial wild beast example illustrates why the substantive evaluation of relevant alternatives under the main principle is not sufficient to determine whether a change or an exception would be an improvement. The reason is that any change can be conceptually divided into two parts: a potential new *substantive* practice and a practice for *implementing* the new

practice. The substantive practice might be ranked very high by the main principle, but the practice for implementing it might be disastrous. The main principle will endorse a change in the status quo only when the combined evaluation of the new substantive practice and the practice for implementing it ranks higher than the status quo and at least as high as any of the relevant alternatives.

To see that there is an issue of implementation of the equal chance of selection practice, imagine that each of the four of us had decided independently on our own version of an equal chance of selection practice and then each of the four of us had used our own random coin-tossing device to select the person who should exit the cave first. Now we have four different instances of a practice that, when substantively evaluated by the main principle, would be ranked above the libertarian status quo and would be ranked equal to the others. But if we implement them all, the result is chaos. Suppose that I am selected to go first by two of the four procedures, and that you and one other member of the group are each selected to go first by only one of the two other random procedures. Then it might be thought that another procedure could be brought into play: The person to go first is the one, if any, selected by more of the four procedures than anyone else. But why that rule? Why not select the last member of the group, the one who was *not* selected by *any* of the four procedures? Or why not keep running the procedures until on some trial, all four pick the same person?

We seem to have uncovered a potential infinity of different practices that are all equivalent when substantively evaluated by the main principle, because a substantive evaluation presupposes successful implementation. This simple example is a reminder that shared moral practices are like laws in that they typically solve a *coordination problem*.[3] Shared moral or legal norms serve to coordinate expectations and behavior, and this coordination often plays a large role in promoting people's life prospects. For a change in a moral or legal practice to be favored by the main principle, it is not enough to consider only a substantive evaluation of the practice itself. It is also necessary to evaluate the practice of implementation. The evaluation under the main principle of any potential change to an existing moral practice or legal practice will typically involve a combination of both kinds of evaluation, a substantive evaluation of the practice, assuming successful implementation, and an evaluation of the practice of implementation itself. Although the combined evaluation contains both elements, it is often useful analytically to discuss them separately, because in a typical case, the question will be whether the expected benefits of a new practice, once implemented, are enough to outweigh the expected costs of implementation (where costs and benefits are evaluated not in dollars, but in effects on life prospects). Typically, the costs of implementation will be the costs of solving the corresponding coordination problem.

This coordination problem can be a very serious problem in many contexts. It will turn out that there is one kind of solution to it that will often be

favored by the main principle: to use a procedure agreed to by a majority. An implementation practice based on majority consent is not by itself enough to guarantee that a substantive practice favored by a majority is endorsed by the main principle. Perhaps the practice is one whereby a majority exploits a minority. That would not be endorsed by the main principle. But when a *substantive* practice is endorsed by the main principle, the *implementation* practice of majority consent will typically produce a combined practice that will be endorsed by main principle.

Have I been too quick to settle on majority consent as an implementation practice? Clearly any supermajority rule would also solve the coordination problem. Why not require unanimous consent? When evaluated by the main principle, the problem with unanimous consent (or any kind of supermajority consent) as an implementation practice is that it is more difficult to obtain than majority consent, so the costs of the unanimous consent (or any kind of supermajority consent) as an implementation practice would be greater than the costs of majority consent. These costs, of course, include the fact that there would be some cases in which worthwhile changes would be made if only majority consent were required, but would not be made for lack of unanimous (or supermajority) consent. When evaluated under the main principle, majority consent is generally favored over any supermajority consent practice, because majority consent is in general the minimum level of consent necessary to avoid the coordination problem and thus is the implementation practice that solves the coordination problem at the least cost. However, as the size of a group increases, the costs of obtaining majority consent can become so high that other kinds of implementation practices can be favored, as I illustrate shortly.

As I described the equal chance of selection practice above, it included the provision that the random coin-tossing procedure was agreed to by a majority. This provision was added so that the implementation practice would solve the coordination problem discussed above. Thus, the equal chance of selection practice is endorsed by the main principle over both the unequal chance of selection practice and the status quo, when evaluated as a substantive practice as a practice of implementation.

Notice that this is not necessarily an argument that the main principle favors majority rule in the political realm. In this example, the goal is to find an implementation practice that singles out one from a potential infinity of *equivalent* practices, all having the same ranking, when substantively evaluated under the main principle. In the political realm, majority rule is used to implement one law from among a group of relevant alternatives that are typically *not* ranked the same by the main principle. Nothing I have said so far implies that the main principle would favor a democratic decision rule in the political realm. However, because laws solve coordination problems, it is at least a point in favor of majority rule in the political realm that it can solve this kind of coordination problem. Of course, if democracies enacted bad laws that greatly impaired well-being when successfully implemented, there

might be no advantage to solving the coordination problem, when evaluated by the main principle. When expectations and behavior are coordinated to efficiently produce harm and misery, lack of coordination can seem positively benign. So it is a separate question, which I discuss in chapter 10, whether the main principle endorses some sort of democratic political system.

It should also be noted that I was implicitly assuming that there was not already an established convention among the four of us for implementing changes to the status quo practices endorsed by the main principle. Perhaps there is a convention that the oldest member of the group is to choose randomization procedures. If I am the oldest member of our group, then this convention would designate my coin toss as the one that binds the group.

I do not think that conventions of this kind represent alternatives to majority consent for solving the coordination problem for implementation practices, because for something to be a convention everyone (or almost everyone) must accept it. So conventions are just one way that the procedural coordination problem can be solved by majority consent.

Equity, Not Necessarily Equality

There is a tendency to think of equity considerations as favoring equality. As the previous example illustrates, often they do. In that example, the fact that the first alternative gave everyone an equal chance of being selected and thus an equal chance of surviving made it more equitable than the second alternative. But the main principle does not always favor equality. Equitably promoting life prospects sometimes requires more rather than less inequality, as illustrated by the following example:

Wild Beast Plus Quicksand Example. Suppose that to get home safely requires not only avoiding being eaten by the beast but also passing through an area with treacherous quicksand. I am the only one of the four of us with the expertise to be able to determine where walking through the quicksand is safe and where it is not. By yourselves, the rest of you would have practically no chance of successfully getting through the quicksand. With me along, everyone would be sure to get through. In this more complex scenario, consider your chances of survival under the same two procedures:

Equal Chance of Selection Practice. This is the practice in which a majority agrees on a fair, random process that gives each of us an equal chance of being selected to be the initial escapee. Then this random process is employed to determine who must exit the cave first. What is your chance of survival under this practice? It depends on who is selected to go first, but the calculation is a little more complicated than in the previous example. Here are the relevant possibilities: (a) I am selected to be the initial escapee. The probability of this happening is 1/4. If it happens, all four of us have a .05 probability of surviving, because none of you will be able to survive the

quicksand unless I survive to lead you through it. (b) You are selected to be the initial escapee. The probability of this happening is 1/4. If it happens, your chances of surviving the wild beast are 1/20 (.05). If you survive the wild beast, I will guide you through the quicksand. So your chances of getting out alive are .05. (c) Someone other than you or me is selected. The probability of this happening is 1/2. If it happens, you and I are both sure to survive.

The calculation of the chances of surviving under the equal chance of selection practice is the same for everyone except me. The probability that any of the three of you will survive under the equal chance of selection practice is .525 (= .25(.05) + .25(.05) + .5(1)). For me, on the other hand, because the quicksand is not a threat to my survival, the calculation of my probability of surviving on the equal chance of selection practice is the same as in the preceding example. My probability of surviving is .7625.

Unequal Chance of Selection Practice. This is the practice in which a majority agrees on a fair, random process that gives each of the three of you a 1/3 chance of being selected to be the initial escapee and I am guaranteed not to be selected. This procedure gives each of us the same probability of surviving as the unequal chance of selection practice gave us in the initial wild beast example: for me, the probability is 1.0; for each of the three of you, the probability is .6833.

In this case, when substantively evaluated by the main principle, the unequal chance of selection procedure is superior to the equal chance of selection practice. This is easily seen, because at the time of the coin toss, all four of us have a higher probability of surviving under the unequal chance of selection practice than under the equal chance of selection practice. Of course, under both practices, I have a higher chance of surviving than any of the three of you. However, under the second practice, the difference between my chance of surviving and yours (1.0—.6833 = .3167) is *greater* than that difference under the first practice (.7625—.525 = .24) and, yet, still the main principle favors the second practice. So the main principle can favor more inequality in life prospects over less.

In this simple case, the alternative favored by the main principle is the same as the alternative favored by maximin—that is, the rule of maximizing the minimum of life prospects. The minimum life prospects under the unequal chance of selection process (.6833) is higher than the minimum life prospects under either the equal chance of selection process (.525) or the status quo (.05). In the next chapter I explain why the main principle does not always agree with the maximin rule.

What about implementation? The unequal chance of selection practice includes a requirement of majority agreement on the fair random process employed, so as to solve the coordination problem. Because these two practices are the only relevant alternatives to the status quo, when evaluated as a substantive practice and an implementation practice, the unequal chance of selection practice is endorsed as an exception to the status quo by the main principle.

Status Quo Shared Background Expectations

The variations on the wild beast example are a reminder that, in the discussion of examples, it is almost always possible to add details that change the moral evaluation of the example. One category of morally relevant factors is particularly important in the present context: the reasonable expectations of those involved. Sometimes those expectations are the result of prior acts. For example, the analysis of the wild beast examples would be very different if it were true that you had induced the other three of us to explore the wild beast's cave by promising to go first if we were trapped in the cave by the beast.

Often expectations are produced not by explicit promises but by the framework for reasonable expectations produced by legal system or customs or ways of life. For example, in many cultures, it would be expected that if there were both men and women in the group, the choice of the initial escapee would be made from the men in the group; women would be exempt. Or in some cultures the class or caste status of the four captives would be relevant to the selection of a procedure to determine who should go first. Call these expectations the *background expectations*.

One of the most important roles of laws, customs, and ways of life is to produce *shared background expectations* about the normative attitudes and behavior of others, because these shared expectations function to solve coordination problems in ways that promote life prospects. These shared background expectations are so important for societies to equitably promote the life prospects of their members that it is unusual for a unilateral change to a shared system of laws, customs, or ways of life to be endorsed by the main principle, even if a substantive evaluation of the practice under the main principle favors the change. The main principle will typically endorse changes in the existing social practices only when they can be and are implemented in a way that produces uniform changes in the shared background expectations—that is, when the changes can be implemented smoothly without significant disruption.

Hart refers to the conventions for recognizing laws as *rules of recognition* (1961, 92). If they are to perform a coordinating function, all systems of norms require rules of recognition, so I apply the term more broadly than Hart does.

Because the production of shared background expectations is such an important part of the way that laws, customs, and ways of life equitably promote life prospects, the main principle establishes a strong presumption in favor of the status quo. This presumption translates into a prima facie duty to obey the law, even if it is flawed, and a prima facie duty to conform to existing moral norms, even if they are flawed. However, not all kinds of exceptions generate coordination problems. One important kind of exception that typically does not generate a coordination problem is an unconscionability exception, which I discuss more fully in the next chapter.

The Example of Privately Enforced Solutions to Collective Action Problems (CAPs)

One of the primary applications of the main principle is to legal solutions to CAPs. Such laws represent a very large exception to the libertarian prohibition on coercion. Nozick has raised a theoretical problem for principles, such as the main principle, that endorse coercive solutions to CAPs (1974, 93–94). If Nozick's argument were correct, the main principle would have to be rejected. So it is important to see that the problem raised by Nozick is not a problem for the main principle.

Music Lovers' CAP. Suppose, for example, that you have a neighbor, Tom, who knows what kind of music you like. One day Tom starts playing the kind of music you like over an outdoor loudspeaker at times when you are home to enjoy it. You do enjoy it. Then one day Tom comes to your door and orders you to sign up for a regular shift as neighborhood disc jockey. When you protest, Tom invokes the main principle to explain why his coercion is part of a practice that equitably promotes life prospects. Only people like you, who enjoy the music, are compelled to serve a shift as neighborhood disc jockey, so all those who are coerced benefit from the practice that includes the coercion.

Is the situation truly a CAP? For it to be a CAP, it is not enough that you and the other potential disc jockeys enjoy the music. You must enjoy it enough that you think your life with the music would be better than your life without it even if you had to take your turn working as DJ.[4] How could Tom know this about you? The best way would be to obtain your consent to being a disc jockey. If you consented, that would indicate that you thought the arrangement enhanced your life prospects. This illustrates why, if it is understood noncategorically, the actual consent exception to libertarian natural rights would be endorsed by the main principle. But what if you do not consent? To make the best case for Tom's coercive practice, let's suppose that the practice of forcing people like you to serve as a DJ for his music would equitably promote the life prospects of you and the other people forced into service. In addition, let's suppose that Tom has worked out a schedule of DJ assignments that fairly distributes the benefits and burdens to all those who enjoy the music he has been playing. And, finally, to avoid other complications, let's suppose that everyone within earshot of Tom's speakers likes the kind of music that you and Tom like, so that the life prospects of everyone affected by Tom's practice are enhanced. Let's say that Tom has found a fair solution to the *music lovers' CAP*.

Is this enough for the main principle to endorse Tom's coercive solution to the music lovers' CAP? I begin with the substantive evaluation. By hypothesis, Tom's solution is ranked above the status quo by the main principle. However, this does not ensure that the practice passes the main principle's substantive evaluation test, because that test depends on what the other relevant alternatives are. When Tom's practice of private enforcement of fair

solutions to the music lovers' CAP is compared with the relevant alternatives, it comes up short, even at the stage of substantive evaluation. There are many systems for producing music that would better promote people's life prospects than the practices of forcing people to serve as DJs for music stations that they enjoy listening to. Obvious alternatives would be to finance the station by advertisements or by subscription. Almost everyone would prefer to listen to ads or to pay their fair share of the cost of music stations they like to listen to rather than to be conscripted to serve as a DJ, and almost everyone would rather listen to stations with professional DJs than to stations staffed by coerced amateurs.

But there is a deeper problem with Nozick's example, which has to do with Tom's practice of implementing his solution to the music lovers' CAP. It generates a coordination problem. Just as in the initial wild beast example, in which the very large number of fair procedures for implementing the relevant practices (e.g., the equal chance of selection practice) generated a coordination problem, here the large number of potential fair solutions to CAPs generates a corresponding coordination problem. To see the problem, imagine that while Tom is scheduling your shift playing music on his stereo, another neighbor, Paul, is scheduling your shift playing music on his stereo, and yet another neighbor, Ralph, is scheduling your shift playing music on his stereo. Suppose it is true that any one of their music projects would be a fair solution to a CAP. Nonetheless, the combination of all three music projects would drastically reduce your life prospects, not only because you would be spending so much time working shifts as a disc jockey, but also because the resulting cacophony of music would be intolerable. And that is to consider only your neighbors who are planning music projects.

When you consider the almost endless variety of ways in which someone could enforce a fair solution to a collective action problem on you, the prospects are nightmarish. This is another example of the coordination problem. In the example of the wild beast, one solution to the coordination problem was to require that a majority agree on the appropriate instantiation of the relevant practice. It might seem that a similar move could solve the coordination problem here. However, there is no such simple explanation of the appropriateness of majoritarian solutions to CAPs. Establishing the institutions of a democratic government requires solutions to lots of CAPs and lots of coordination problems. I return to this topic in chapter 10.

There is another way of solving the coordination problem illustrated by the music lovers' CAP. Everyone could get together and agree to confer on one person the monopoly power to enforce solutions to CAPs, including the power to force everyone else not to enforce coercive solutions to CAPs. This is the idea behind Hobbes's [1651] defense of an absolute sovereign. Because of the constraints of his theory, Hobbes thought that he had to show that everyone would consent to granting the sovereign such powers in order for the powers to be legitimate.[5] But the main principle does not imply any such constraints. It would seem that the main principle would endorse the practice

of having an absolute sovereign to enforce solutions to CAPs, including enforcing a prohibition on anyone else from enforcing coercive solutions to CAPs. It turns out that the main principle is not so supportive of absolute sovereigns, but I put off the discussion of that topic to chapter 10. In the interim, it simplifies the discussion to assume that there is some acceptable implementation practice for the changes in moral and legal practices that I discuss. This enables us to focus on the substantive evaluation of them.

What are we to say about Nozick's example of the music lovers' CAP? It fails as a counterexample to the main principle on multiple grounds. In fact, there is a certain irony to this example. Although Nozick used the example as part of an argument against consequentialist views or other views that allow coercive solutions to CAPs, he never said exactly why Tom's solution to the music lovers' CAP is not justifiable. The irony here is that one potential explanation of why Tom's use of coercion to solve the music lovers' CAP is not justifiable is that we would all be *better off* if most private enforcement of CAPs were prohibited[6]—that is, that prohibiting most private enforcement of CAPs is itself the solution to a CAP. Prohibiting private enforcement of solutions to CAPs would equitably promote life prospects. If this is right, then not only is Nozick's example *not* a counterexample to the main principle, but the main principle can explain the result that Nozick reaches. Of course, Nozick would not agree that the main principle is the *correct* explanation of the result he reaches in this example. But it is to the explanatory credit of the main principle that it does explain that result—that is, that it explains why it would not be morally appropriate for Tom to make an exception to the status quo prohibition on coercion to use coercion to enforce his solution to the music lovers' CAP.

The Wilt Chamberlain Example

Nozick raises another challenge to consequentialist theories. He claims that no consequentialist theory can adequately represent the moral significance of liberty, because consequentialist theories are patterned, but "liberty upsets patterns" (1974, 160). He offers what he regards as a potential counterexample to all consequentialist principles (1974, 160–164).

Nozick invites us to consider a social arrangement satisfying our favorite pattern in distributing well-being (in the case of the main principle, equitably promoting life prospects). Suppose that in that ideal social arrangement there are millions of people who are willing to trade some of their resources to a basketball player named Wilt if they can watch him play basketball. Wilt's income from playing basketball will give him lots more resources than other people, and lots more than he had under our favorite pattern. If the initial arrangement was equitable, it would seem that the final one, in which Wilt has a much larger share of social resources, would not be. But everyone involved prefers to have made the trade than not to have made it. Should

they be prevented from making the trade, simply to preserve our favorite pattern (e.g., equity)? Nozick claims this example shows, in his memorable words, that those who would preserve their favorite pattern would have to "forbid capitalist acts between consenting adults" (1974, 163).

Technically, it would not be necessary to *forbid* capitalist acts between consenting adults. The same result could be achieved by undoing them. Suppose, for example, that there were a bureau of equity that would intervene to confiscate resources and redistribute them, whenever doing so would promote equity. The bureau would not forbid Wilt from charging people to watch him play basketball, but it would remove his incentive to play if it were generally known that if people did pay him, the bureau would intervene to confiscate all or almost all of his income and to redistribute it according to the requirements of equity. If he could not keep his basketball income, Wilt would not play and everyone involved, Wilt and his fans, would be worse off.

It should be surprising that Nozick claims that this example shows how liberty upsets patterns, when the example itself is an example of how coercive enforcement of a pattern (e.g., equity) can make everyone involved worse off in their own estimation. Making everybody worse off is itself a pattern and one that is not endorsed by most consequentialist principles. So it would seem that there is a consequentialist explanation of why coercive enforcement of patterns is morally inappropriate.

What Nozick's example shows is that the *practice* of forbidding capitalist acts between consenting adults or a practice that has the same effect (e.g., the bureau of equity) would *not* be endorsed by the main principle, because it would be reasonable to expect it to reduce the life prospects of everyone (or almost everyone) affected. As a general rule, the main principle endorses practices that *promote* people's life prospects, not practices that *reduce* them. Thus, as in the example of the music lovers' CAP, it turns out that the main principle can *explain* the result that Nozick reaches in his example. So this example does not cast doubt on the main principle, it supports it.

What Nozick's example really illustrates is *not* that no patterned principle can be acceptable, but that an acceptable principle must apply to practices, not to individual acts. The bureau of equity illustrates the fact that a requirement that every individual economic transaction promote equity would have disastrous results from the point of view of equitably promoting everyone's life prospects. Evaluated as a practice, such a requirement would almost surely make everyone worse off than a practice of permitting voluntary economic transactions (capitalist acts). Thus, the correct moral to draw from Nozick's example is that an acceptable consequentialist principle should evaluate practices, not individual acts.

Nozick's discussion also shows that once attention is focused on practices, the practice of permitting unrestricted capitalist acts could be better for everyone than the practice of completely forbidding them. This, by itself, is not enough to show that the main principle would endorse the practice of

permitting unrestricted capitalist acts, because there might be *other* practices that would do a better job of equitably promoting people's life prospects than either of the policies that Nozick considers. Suppose it were possible to establish a progressive income tax, requiring those with higher incomes to pay a higher tax rate, but the rate would not be so high as to take away the motivation of Wilt and others with high incomes to do the things that would earn them those high incomes. Then Wilt would get the benefits of being able to keep some percentage of his basketball income, the fans would get the benefit of being able to watch him play basketball, and the taxes from his basketball income could be used to promote equity—for example, by providing educational opportunities for those who could otherwise not afford them. This sort of social practice would better promote equity than a practice that allowed Wilt to keep all of his basketball income, so it would be favored by the main principle's substantive evaluation test. So long as the costs of implementation were not too great, a practice of progressive taxation would be ranked above both the practice of forbidding capitalist acts between consenting adults and the practice of unrestricted capitalism.

Once it is recognized that the main principle evaluates practices rather than individual acts, the application of the main principle to the Wilt example shows that it *can* explain the reasons for promoting liberty and the reasons for not forbidding capitalist acts between consenting adults, though it does not necessarily support the practice of unrestricted capitalism. I discuss economic rights in chapter 9.

A Bureau of Equity?

The example of the bureau of equity illustrates a general problem with applying the main principle. It might seem that it is the lack of a precise formula for the equitable promotion of life prospects that is the greatest impediment to morally improving the world. This would be a mistake. Even if we had such a formula, most of the difficult issues about how to equitably promote life prospects would still have to be solved. To see why, suppose that there were a generally accepted formula for ranking social practices in terms of how well they equitably promote life prospects. Call it the *equity formula*. One idea for promoting distributive justice would be this: Establish a bureau of equity, and empower the bureau to confiscate resources of the well off and redistribute them to the less well off whenever doing so would better equitably promote life prospects according to the equity formula.

The problem with this suggestion is that even if each individual act of confiscation and redistribution promoted equity according to the equity formula, the indirect effects of such a social practice could well be disastrous for everyone. The confiscatory actions of the bureau of equity might well eliminate or greatly reduce incentives to make worthwhile investments. As a result of these indirect effects, everyone's life prospects might well be

reduced, which would not be a way of equitably improving them. This is another reminder of the importance of evaluating acts as social practices or as embedded in social practices.

The present argument is not an argument that no policies of redistribution could ever be justified, but only that the social practice of making individual confiscations and redistributions whenever they would directly promote equity would almost surely be self-defeating. Thus, even if the bureau of equity had a precise formula for equity, their most difficult problem would be to design a *policy* that would satisfy the formula when both the direct and indirect effects of the policy, as applied by human beings as we know or are justified in believing them to be, were taken into consideration. This is not an easy problem to solve.

Initial Statement of the Main Principle

In this chapter we have learned a lot about the main principle. It evaluates changes or exceptions to status quo practices. Proposals for improvements are not evaluated in some ideal world of strict compliance or ideal justice, but in the actual world given by the background of status quo practices. At this point I discuss moral and legal practices separately. Ultimately, the main principle will show us how to unify the discussion of the two practices.

The main principle's core ranking of practices is the basis for its evaluation of potential changes or exceptions to the status quo. This evaluation can be analytically divided into two parts: the substantive evaluation of the new practice and the evaluation of the practice of implementation. Typically the new practice must have enough substantive benefits to outweigh the costs of implementation, which are usually the costs of solving a coordination problem. However, as the unconscionability exceptions to ground-level moral norms illustrate, sometimes implementing a new practice does not generate a coordination problem. In such cases, there may be little or no implementation costs.

The examples discussed in this chapter have given us a good idea of how the main principle provides a sufficient condition for a change in moral or legal practices to be an improvement. Let's attempt a preliminary statement of this condition:

Initial Statement of the Main Principle. A change in or exception to status quo moral or legal practices is endorsed as an improvement by the main principle just in case the change, when evaluated as a substantive social practice and as a practice of implementation, would make the overall system of social practices one that does a better job of equitably promoting life prospects than the status quo system of practices and also does a better job than any of the relevant alternatives.

There is one further indeterminacy in this statement that requires clarification. Whose life prospects must be equitably promoted by a change in social practices for it to be endorsed by the main principle? The simplest

answer would be everyone affected by the relevant practices. This answer overlooks important distinctions among those affected by a practice. First, we need to distinguish those affected by the practice into two groups, potential cooperators in the practice and bystanders. If the practice under consideration is a change in a society's legal system, bystanders would be those in other societies who are not bound by that legal system. The only constraint on the life prospects of bystanders is that the practice not *lower* their life prospects. Consider, for example, a law in society S that required everyone in S to dump their garbage in society T. Everyone in society S would be bound by the law, so they would all be participants in the practice. The members of T would be bystanders. Such a law might very well promote the life prospects of everyone in S. However, it would not satisfy the main principle, because it would lower the life prospects of the members of T, who would be bystanders to the practice.

It is only participants in a practice who can reasonably expect that their life prospects should be equitably promoted by it.[7] To explain how the main principle applies to participants in a practice, consider how the main principle would apply to a system of property rights that includes a law against destroying the property of others. We need to consider three categories of participants in the practice:

1. *Compliers*: those who comply with the practice by not destroying the property of others.
2. *Responsible noncompliers*: noncompliers who are morally responsible for their noncompliance. This category includes those who destroy the property of others intentionally, those who act in ways that they foresee will lead to the destruction of property of others though they don't intend it, and those who destroy the property of others due to negligence.
3. *Nonresponsible noncompliers*: noncompliers who are not morally responsible for their noncompliance. This category would include someone who was unlucky enough to have her unoccupied car explode and damage other people's property, where there was no reason for her to have anticipated that her car would explode, or young children who cause property damage to others.

The reason for distinguishing these three categories is to be able to precisely identify the category of participants whose life prospects are typically *not* covered by the main principle—the responsible noncompliers. The responsible noncompliers are those who are liable to punishment for noncompliance. As I explained in chapter 1, the main principle does not apply to enforcement, including punishment. It is easy to see why the main principle does not typically apply to responsible noncompliers. No legal system could ever be justified if justifying it required it to promote the life prospects of those who break the laws. This is not to say that there are no limits on the

treatment of responsible noncompliers, only that those limits are not explained by the main principle. Throughout this book I assume that there are some proportionality constraints on acceptable treatment of responsible noncompliers, because, for example, it could not be permissible to execute those who knowingly double park, even if doing so would drastically reduce the amount of double parking and, thus, increase the life prospects of compliers and nonresponsible noncompliers.

Final Statement of the Main Principle

We can now formulate a more precise statement of the main principle condition:

Final Statement of the Main Principle. A change in or exception to status quo moral or legal practices is endorsed as an improvement by the main principle just in case the change, when evaluated as a substantive social practice and as a practice of implementation, would not reduce the life prospects of bystanders and would make the overall system of social practices one that does a better job of equitably promoting life prospects of all participants, except those covered by the responsible noncompliance exclusion, than the status quo system of practices, and also does a better job than any of the relevant alternatives.[8]

The *responsible noncompliance exclusion* is this: The life prospects of responsible noncompliers are excluded from consideration under the main principle, so long as it is true that *if* they had complied with the practice, their life prospects would have been equitably promoted by the practice.

The *if-they-had-complied condition* in the responsible noncompliance exclusion is meant to rule out cases of the following kind. The monarch Phillip announces a new law requiring Jews to convert to Christianity or to forfeit their property to be distributed equitably among the Christians. Phillip correctly predicts that no Jews will convert. As a result, all Jews forfeit their property and the life prospects of those who comply with the law (all the Christians) are equitably promoted. Suppose there are no nonresponsible noncompliers. This law fails to be endorsed as an improvement by the main principle even though it equitably promotes the life prospects of all participants except responsible noncompliers, because *if* the responsible noncompliers (the Jews) had complied, the law would not have equitably promoted the life prospects of compliers. Having to give up their religion would have greatly diminished the life prospects of the Jews.

A full analysis of a potential change in a status quo practice would require a comparison of the substantive practice and the practice of implementation with the status quo practice and the other relevant alternatives based on the effects on bystanders (that their life prospects are not diminished), by their effects on responsible noncompliers (to determine if the responsible noncompliance exclusion applies), and by their effects on compliers and

nonresponsible noncompliers (to determine if their life prospects are equitably promoted, by comparison to the relevant alternatives). To carry out such a complex analysis in every case would make the discussion of examples in coming chapters dauntingly complex. Fortunately, most of the examples I discuss do not require such a complex analysis. To simplify the discussion, unless I specify otherwise (e.g, in the discussion of systems of strict liability in chapters 5, 6, and 9), I assume that the practices that I discuss are practices that do not adversely affect bystanders; that they are practices to which the responsible noncompliance exclusion applies; and that they are practices that similarly affect the life prospects of compliers and nonresponsible noncompliers. This enables me to limit my discussion to the effects of the relevant practice on the life prospects of those who comply with the practice. Also, I will often set aside questions about the practice of implementation, so that I can focus entirely on the substantive evaluation of the relevant practice under the main principle. In this way I can simplify the exposition with no loss of generality.

The Main Principle as a Principle of Moral Reciprocity

Although the main principle places a constraint on the effects on bystanders (that their life prospects not be reduced), it typically only requires that a practice equitably promote the life prospects of those who comply with the practice (or of those whose failure to comply is not due to their own fault or negligence). Compliers and nonresponsible noncompliers can be characterized as those who are willing to cooperate on the terms given by the practice, given that the others are, also. So I refer to them as *cooperators*. Because the main principle requires that practices equitably promote the life prospects of cooperators, I say that the main principle is the main principle of *moral reciprocity relations*.[9] I interpret these reciprocity relations broadly—so broadly that they can extend backward to past generations and forward to future generations yet unborn. However, for the purposes of this book, I limit my focus to moral reciprocity relations between human beings who are contemporaries of each other.

The main principle provides a sufficient condition for improvement in moral and legal practices. It is not necessary, because not all of morality concerns moral reciprocity relations. Some of our duties to other moral agents—for example, humanitarian duties of assistance—do not depend on moral reciprocity relations, though such duties are typically more stringent when they involve agents who have moral reciprocity relations with each other. Although it is possible to think of our duties to nonhuman animals as moral reciprocity relations in an extended sense, I do not assume that they must be understood that way. Finally, I leave open the question of whether we should think of ecosystems as an independent source of moral obligation.

As I interpret it, the Golden Rule is an attempt to capture moral reciprocity relations in a ground-level moral principle. It is not completely adequate for

the reason that no ground-level moral principle can ever be completely adequate. But once a moral tradition has adopted a Golden Rule principle, it has reached the stage where it has become sensitive to moral reciprocity relations and, thus, where the main principle provides a sufficient condition for moral improvement.

The idea of moral reciprocity relations is the guiding idea behind Rawls's theory of distributive justice. The main principle is my way of extending that idea to most primary ground-level moral practices in moral traditions that have crossed the consequentialist threshold. To equitably promote life prospects, a new social practice does not have to be a Pareto improvement over the status quo. For example, the elimination of an institution of slavery can be a moral improvement even if it reduces the life prospects of the slave owners. But a change in social practices could not be endorsed by the main principle if it improved no one's life prospects, unless it were itself part of a social practice of making changes that, over time, equitably promoted life prospects.

Rawls's proposal for understanding moral reciprocity relations was to think of them as the relations between those in a system of cooperation that fairly distributes the benefits and burdens of cooperation. Nozick accused Rawls of treating the benefits of cooperation as if "they fell from heaven like manna" (1974, 198), rather than as things to which people have entitlements. This criticism is based on a misunderstanding of meta-theoretical explanation. Because Rawls's theory was a meta-level theory of how ground-level entitlements can be justified, of course, it could not be constrained by the ground-level norms of entitlement. Rawls's theory had to abstract away from the existing ground-level norms of entitlement to consider whether they are justified.

But even in abstracting away from the existing ground-level entitlement norms, Rawls did not treat the benefits of cooperation as "manna from heaven." He supposed that those benefits were the joint product of social cooperation, and so were to be fairly divided among the cooperators. This is one consequence of thinking of distributive justice or of ground-level moral practices as capturing moral reciprocity relations, and it distinguishes the main principle from other consequentialist principles.

The main principle does not regard the benefits of a cooperative practice as "manna from heaven" to be distributed in whatever way will do the most (appropriately weighted) good. Consider an example. Suppose that two countries have no shared cooperative social practices and that one is very wealthy and the other very poor. The main principle contains no presumption that the social practices of the wealthy country must equitably promote the life prospects of those in the poor country. So even if there are norms of *allocative justice* (Rawls 1971, 88) that required that manna from heaven be given to those most in need (e.g., those in the poor country), because the main principle is a principle of moral reciprocity (which includes *distributive* justice, as Rawls uses the term), it does not treat the benefits of cooperation as manna

from heaven in this way. This is not to say that a wealthy country should regard itself as having no moral obligations to poor countries with which it has no shared cooperative practices, only that the main principle can explain how those obligations could be different from and less stringent than its obligations to its own poor.[10]

I should mention that, as a principle of moral reciprocity, the scope of the main principle is not as narrow as the scope of Rawls's principles of distributive justice. Rawls was never able to include those with special health care needs or those with severe disabilities in his theory. I propose a way of doing so in chapter 4 and actually do it in chapter 11.

Is the Main Principle a Principle of Justice?

Rawls presented his two principles as principles of justice. Are they principles of justice? I don't believe so. Consider, for example, an egalitarian hunter-gatherer society. Suppose that this society comes into contact with an inegalitarian capitalist society in which everyone's life expectancy, standard of living, and so forth are much higher than in the egalitarian society. If the members of the egalitarian society voluntarily enter the inegalitarian society, the main principle might well endorse the change as an improvement, even though, on the ordinary notion of justice, the egalitarian society is more just than the inegalitarian one. This is why I agree with G. A. Cohen that Rawls's theory is not really a theory of justice (2008, chap. 7).[11] For the same reason, the main principle is not a principle of justice either. It is a principle of moral improvement.

What Kind of Principle is the Main Principle?

What kind of principle is the main principle? First, it is a meta-level explanatory principle, not a ground-level principle to be applied in moral or legal deliberation.

Second, it is an objective principle that determines when a change in moral practices would be an objective improvement, not a principle that evaluates motivation. However, it plays a crucial role in explaining moral motivation, because, as I discuss in chapter 5, to develop moral judgment (as opposed to mere moral rule following) is to develop an implicit sensitivity to the main principle.

Third, it is a comparative principle, not an optimizing principle, because, in any given situation, it does not require optimizing; it simply endorses the best among a limited number of relevant alternatives.

Fourth, the main principle is a *patterned* principle (Nozick 1974, 156), because it specifies a complex pattern of life prospects (equitable distribution of life prospects at every stage of life) by reference to which alternative moral and legal practices are evaluated.

Fifth, it is a partly *historical* principle (Nozick 1974, 153), because history matters. The main principle makes historically based distinctions—for example, it differentiates between bystanders and participants and does not take into consideration the life prospects of responsible noncompliers (when the responsible noncompliance exclusion is satisfied). Their life prospects are covered by other principles of proportionality in enforcement and punishment.

Sixth, when attention is limited to cooperators (i.e., compliers or nonresponsible noncompliers), the main principle considers only changes in the distribution of life prospects over time to determine whether a change in a moral or legal practice is an improvement. In Nozick's terms, that makes the main principle's test a *multiple time-slice end state condition* (Nozick 1974, 155), or even more precisely a *multiple time-slice condition on (population) distributions of (individual) distributions of life prospects*.

Finally, the main principle gives the (appropriately distributed) good *explanatory priority* over the right. Ground-level concepts of *rightness* in morality, *justice* in the law, and even the priority of *human rights* are explained at a deeper in terms of (appropriately distributed) well-being. It is true that one's share in the equitable distribution of well-being must itself be earned by willingness to comply with the relevant moral or legal practices. However, that does not change the fact that the *content* of those practices is determined solely by their *consequences* for the distribution of well-being (evaluated as life prospects) among cooperators. It is this fact that makes the theory a multiple time-slice end-state theory and that makes me conclude that it should be classified as a *consequentialist* theory of moral improvement.[12]

Nozick used the Wilt Chamberlain example to argue that liberty upsets patterns. However, when the focus shifts from individual acts to the practices that include them, the Wilt Chamberlain example *fits* a pattern, a pattern of promoting well-being. So Nozick was mistaken to think that the Wilt Chamberlain example ruled out all patterned principles or all end-state theories. The main principle provides a competing explanation to Nozick's own theory that must be taken seriously.

However, there are a number of other objections that have been thought to be fatal to consequentialist accounts of moral and legal practices. I address many of them in the next chapter and use them as an opportunity to further clarify and motivate the main principle.

FOUR

What Is Well-Being? What Is Equity?

In this chapter, I do my best to more precisely specify the content of the main principle, with particular attention to the concept of well-being, understood in terms of life prospects, and the concept of equity, even though I have no definitions for them. Along the way, I consider some of the reasons that consequentialist accounts of morality and justice are often thought to have been decisively refuted. Responding to these objections will help me to provide a fuller explanation of the important terms in the main principle and of how the main principle avoids the objections that have seemed decisive against other consequentialist views. For simplicity, I limit my discussion to the substantive evaluation of the social practices that I discuss and I set aside questions about implementation.

What Is Well-Being?

Many nonphilosophers think of well-being as a state of feeling really good (e.g., Gilbert 2006). They have a hedonistic theory of well-being. This hedonistic theory probably explains some people's actions—that is, some individuals probably do act to maximize the *net hedonic value* of their lives (i.e., the net sum of pleasure over pain over the course of their lives).

However, most people are not hedonists. Most people have goals other than maximizing the net hedonic value of their lives. When Mill, who espoused a hedonistic theory of well-being, was confronted with examples that indicated that people do not always seek to maximize net hedonic value, he modified the theory to be compatible with this result. Presented with the evidence that he and most educated people would choose the life of a dissatisfied Socrates over the life of a satisfied pig, he introduced a distinction between higher and lower pleasures (i.e., he made net hedonic value a function of both the quality and quantity of one's pleasures and pains) to enable him to claim that the life of a dissatisfied Socrates could have higher net hedonic value than the life of a satisfied pig. If we preferred the life of a dissatisfied Socrates over the life of a satisfied pig, that showed only that even a small amount of the higher pleasures of the life of a dissatisfied Socrates would outweigh a large amount of the lower pleasures in the life of a satisfied pig.

I think that if Mill hadn't been so keen to save his theory from these potential counterexamples, he would have realized that he should have been suspicious

of his introduction of a distinction between higher and lower pleasures into the theory. To introduce a technical term that I explain shortly, the distinction seems to be a *fudge factor* introduced to insulate the theory from counterexamples.

Except to save his theory, why would he even suppose that there are higher and lower pleasures? Why would he think that the pleasures that he ranks as "lower" have less net hedonic value? If anything, it would seem that the opposite is true. Only someone trapped by a theory would think that engaging in philosophical dialogue had greater net hedonic value than stimulation to orgasm. In any case, Nozick (1974, 42–45) has provided us with a fairly decisive objection to any kind of hedonism as a descriptive theory of human motivation. Imagine that you have the opportunity to hook up to a virtual reality experience machine for an hour and have any kind of experience that you desire. During that hour you will really believe that what you are experiencing is real and your pleasures and pains will have all the intensity of pleasures and pains that you feel when not hooked up to the machine. Your assignment is to maximize the net hedonic value of your hour on the machine. Would you choose to have the experience of engaging in a philosophical dialogue or would you choose an intense sexual or gustatory experience? It is hard for me to believe that a majority of those who are familiar with all three kinds of experiences would choose the philosophical dialogue. If so, Mill's claim about the hedonic weight of "higher" and "lower" pleasures is just false. If anything, the "lower" pleasures rank more highly in the hedonic calculus than the "higher" ones.

Even if I am wrong about this, Nozick's experience machine thought experiment can be used to show that Mill's explanation of our preference for the life of a satisfied pig over the life of dissatisfied Socrates is mistaken. Imagine an experience machine on which you could have the virtual life of a *satisfied* Socrates. Instead of being put to death, you are given a teaching award and revered as a model citizen. Your philosophical theory of justice presented in *Republic* is so compelling that you are invited by your fellow Athenians to redesign the city government to fit your model. And the redesign produces a harmonious and flourishing city. (Remember, it doesn't have to work in the real world, just seem to work in the virtual world of the experience machine.) It would be easy to fill out the details of the life in such a way that virtual Socrates would achieve much more of the higher pleasures than the actual Socrates ever did. On Mill's account, the virtual life of satisfied Socrates on the machine would be a happier life than the actual life of the dissatisfied Socrates and, thus, Mill's theory implies that anyone familiar with the kinds of pleasures involved would choose the virtual life of satisfied Socrates over his actual life. This is not true. I am quite confident that Socrates would not have chosen it. I know I would not choose it. Would you choose it?[1]

Mill was committed to the view that happiness could be defined in terms of net hedonic value. I think he was mistaken about happiness. But I admit

that happiness can be given a hedonic interpretation. That is the reason that I have focused on well-being rather than happiness. I want a term that is not so likely to be understood in hedonic terms.

There is another reason for thinking that net hedonic value is a poor proxy for well-being, the phenomenon of adaptive preferences (e.g., Nussbaum 2000, 136–142; Sen 1999, 62–63). People who live in states of absolute deprivation with no hope of escape adjust their aspirations accordingly. Subjective measures of satisfaction can show them to be more satisfied than others in different circumstances who are much better off on any objective standard. The phenomenon of adaptive preferences shows that, even if well-being has some subjective elements, it cannot be adequately defined solely in terms of subjective factors.

There are many different accounts of well-being in the philosophical literature (e.g., Parfit 1984; Griffin, 1988). The great variety of views is itself evidence that we don't have anything like a definition of well-being. On my view, this is one of the bases of human rights. We need to be free to conduct experiments in living not only to help to determine how to achieve well-being, but also to help to determine what it is. It seems to me that, for human beings, well-being requires at least some worthwhile human relationships and at least some success in achieving worthwhile goals. Neither of those can be achieved in a virtual life on an experience machine, so such a life ranks low on the scale of well-being.

Fortunately, it is possible to develop a theory of moral improvement without a precise conception of well-being, so long as we can make comparative judgments about particular cases. Thus, for example, on any reasonable conception of well-being, even a hedonic conception, to be freed from a system of slavery in which, as a child, you were separated from your parents and sold at auction and, as an adult, your children are separated from you and sold at auction and instead to be allowed to be raised by loving parents and to become one yourself represents a substantial increase in well-being.

Alternatives to Well-Being in Political Philosophy

Over the course of the past 40 years, there has emerged a powerful movement in political philosophy against using the concept of well-being to evaluate laws or political systems. The movement began with John Rawls's proposal that the basic political institutions be evaluated on the basis of the distribution of *primary goods* rather than actual well-being. Rawls's idea was that rather than thinking of political institutions as justified by promoting the well-being of those who live under them, political institutions would be justified as enabling people to form and pursue their own life plans, and as enabling them to obtain a fair share of the goods that would be useful to them in forming and pursuing their life plans, whatever they might be. Rawls classified primary goods into four categories, rights and liberties, opportunities and powers,

income and wealth, and the social bases of self-respect (1971, 92; 1993, 308–309). Rawls thought that it would be possible to specify the conditions for distributive justice in terms of expectations of primary goods rather than in terms of well-being.[2]

What is striking about Rawls's list is that it is almost wholly external. It defines a person's status in terms of her external circumstances, not in terms of her own internal capacities and abilities. This makes it an example of a *resource-based account*. My criticisms apply to almost any resource-based account.[3]

Nussbaum and Sen have proposed alternative accounts that are similar, though not identical. Their accounts focus not on external conditions alone, but on the actual capabilities that a person has, capabilities that are a product of external conditions (e.g., rights and liberties) and a person's own internal physical and psychological capacities. Though there are some differences between Nussbaum and Sen, here I focus on what they have in common. Sen identifies the relevant capabilities as "substantive freedoms . . . to choose a life that one has reason to value" (1999, 74).[4] Nussbaum provides a detailed list of central human capabilities (2000, 71, 78–80).

Consider an example. People tend to think of a right against starvation as a welfare right to resources—for example, as a right to food or to the money needed to buy food. On a capabilities account, a right against starvation might be thought of as a right to what is necessary to be able to earn a living. This will include education and other things that require resources. But it will not necessarily be understood as a welfare right.

As applied to physically and psychologically normal human beings, which is the intended application of Rawls's theory, there is reason to wonder whether there is any significant difference between an account of justice based on Rawls's primary goods and an account based on Nussbaum or Sen's capabilities. The reason is that Rawls could argue that assuring fair equality of opportunity and fair expectations of income and wealth, for example, will require that a society assure that children receive what is necessary to develop their natural talents and abilities. But even in this case, it is useful to augment Rawls's account to explicitly acknowledge the good of developing one's talents and abilities. Rights to subsistence, health care, and education, which are mentioned only in passing by Rawls, have much greater prominence in a theory that focuses on capabilities.

Even more important, Nussbaum and Sen's capability alternatives point to the possibility of extending the theory of justice to apply to those who lack normal human physical or psychological capacities. This would be an important extension of Rawls's theory.

For these reasons, I regard the Nussbaum and Sen capabilities approaches as an improvement on resource-based accounts. But I think it is a mistake to try to defend a capabilities approach on any other grounds than as a proxy for well-being. Indeed, I believe it is an important part of a consequentialist political philosophy to explain why requiring a government to

justify its policies by their contributions to capabilities (rather than directly by their contribution to well-being) is generally the best way of promoting well-being.

To see why an account in terms of resources or capabilities needs to be grounded in well-being, consider how Plato might try to defend the rule of the philosopher-autocrat in his ideal state, the *Republic*, which I refer to as the *beehive society*. The beehive society is one in which ordinary citizens have no substantive freedom to do anything other than what the philosopher-autocrat tells them to do. In the beehive society, citizens develop only the talents and capacities that will enable them to perform their social functions well (e.g., as cobbler), and to be obedient to their superiors. Plato would have regarded almost everything on Rawls's list of primary goods as *bad* for ordinary citizens, including rights and liberties, opportunities and powers, income and wealth. Even the social bases of self-respect would be denied to most citizens, who would be deceived into believing that they were inherently of low caste (Republic III 414b7–415e4). Thus, any account of the justification of political institutions based on how well they provide Rawls's primary goods would rule out *a priori* Plato's beehive society.

Similarly, in Plato's beehive society, most citizens would have only quite limited *capabilities*, for, contra Sen, Plato would claim that it is bad for people to have the substantive freedom to do anything other than what their leaders order them to do; and, contra Nussbaum, Plato would argue that many of the capabilities on her list are bad for people and that they are better off not developing them. Call this *Plato's challenge*.

I do not see how to answer Plato's challenge *a priori*. It is possible for human beings to have been so constituted that having choices would give rise to debilitating anxiety and that having to form their own life plans would be felt as an intolerable burden. Or it is possible for human beings to have had such poor judgment that when they were given an opportunity to form their own life plan, they almost invariably would come up with one that would make them miserable. These are not mere logical possibilities. Plato and other apologists for autocracy typically claim that some or all of them are true. And there is at least some evidence that supports them. So it seems to me that Plato's challenge cannot be ruled out *a priori*.

What's So Good about Resources or Capabilities?

In order to rule out a beehive society, I believe that Rawls, Sen, and Nussbaum would have to argue that it is better for people to develop and exercise their own judgment about what sort of life plan to pursue and how to pursue it than for people to be trained to obey authorities who simply assign each person a life plan. The early Rawls claimed that primary goods would be useful in pursuing any rational life plan (1971, 92, 397, 407–408). He simply

failed to address the possibility that it might not be good for human beings to be allowed to develop their own life plan. In his later work, Rawls explicitly disavowed the metaphysical goal of articulating the ideal form of society for human beings. He limited himself to the political, not metaphysical, project of articulating the "bases of agreement implicit in the public culture of a democratic society" (1993, 339).

Nussbaum does not disavow the metaphysical project, with its universalist aspirations. She makes a good case for the importance of what she refers to as *combined capabilities* even to poor women in the developing countries (2000, chap. 1). But, as addressed to the advocate of the beehive society, there is a crucial gap in her argument.

Nussbaum draws attention to the difference between functioning in "a truly human way, not a merely animal way" (2000, 72). She insists that "[a] life that is really human is one that is shaped throughout by these human powers of practical reason and sociability" (72). The linchpin of her defense of her combined capability account is "the very great importance the approach attaches to practical reason, as a good that suffuses all the other functions, making them human rather than animal" (87). But, of course, to the advocates of the beehive society, one of its benefits is that at least most of its members do *not* have to exercise the power of practical reason. By making human (in her sense) as opposed to merely animal functioning the standard of what is good for human beings, she introduces a factor that rules out the beehive society *a priori*. But whether people would be better off in a society that promotes Nussbaum's combined capabilities or in a beehive society cannot be settled *a priori*.

Sen provides surprising evidence of the ways that the substantive freedoms (with which he identifies capabilities) promote well-being. For example, he has reported that democracies with an active opposition and a free press do not have famines (1999, 178–184). He has also provided compelling evidence that education and opportunities for employment outside the home for women increase their life expectancy and are more effective methods of population control than coercive family limitation laws (chap. 8). But on the question of whether a society that promotes Sen's substantive freedoms is better for human beings, all things considered, than life in a beehive society, he is reduced to quoting Cowper's couplet: "Freedom has a thousand charms to show / That slaves, howe'er contented, never know" (1999, 298).

It is hard to know what to make of this. Even if it is true, it fails to address the possibility that freedom might bring with it a thousand frustrations and anxieties that would more than counterbalance its charms. Whether the frustrations and anxieties of freedom counterbalance its charms is a crucial question in justifying policies that promote Sen's capabilities.

In fairness to Nussbaum and Sen, they are not consequentialists, so they are not required to respond to Plato's challenge. It is only for a consequentialist like me that a full defense of the role of capabilities in moral and political

thought depends on their standing as proxies for well-being. So it is up to me to try to respond to Plato's challenge.

Because Mill was a consequentialist, he also had to respond to it. It turns out that Mill provided us with an example of how not to respond to Plato's challenge before he provided an example of how to reply to it.

Mill's Two-Pronged Response to Plato's Challenge

In *On Liberty*, Mill did not explicitly address the beehive society of Plato's *Republic*, but Mill definitely intended to be addressing and settling the question of whether an experimental society would produce greater overall well-being than a beehive society. He gave two arguments for favoring the experimental society over the beehive society. The first was an application of his distinction between higher and lower pleasures. This argument is very unconvincing. Fortunately, he followed it up with a much more promising one. His second argument will be the model for my consequentialist account of human rights, though my account differs from Mill's in many significant ways—not least, in that it is not utilitarian.

Mill's first argument was to appeal to his distinction between higher and lower pleasures to argue that autonomy rights make possible a life of individuality with the associated higher pleasures that are not even possible for those who live lives of conformity in a beehive society. In one excess of hyperbole, he even claimed that individuality was "one of the leading essentials of well-being" ([1859], 65).

What could have led him to make such an extravagant claim? I think that Mill was trapped by his model of human psychology. Unable to imagine that people's preferences could be based on anything other than hedonic value, he needed to modify his theory of hedonic value to fit people's preferences. So he did. He introduced a *fudge factor* that could be guaranteed to fit people's preferences in the actual world. The fudge factor was the weights assigned to higher and lower pleasures.

A *fudge factor* is a factor introduced into a theory that seems designed to insulate the theory from being falsified. For example, when someone who claims to have the power to bend spoons adds that due to a "shyness factor," the power is ineffective in the presence of skeptics, the shyness factor operates as a fudge factor to his theory that makes it nearly impossible to refute. Note that theories that contain fudge factors are not necessarily false. It is *possible* that the power to bend spoons would be compromised by the presence of skeptics. But we should be suspicious whenever a fudge factor is introduced to save a theory from refutation.

It is not always easy to identify a fudge factor. After all, in my bottom-up methodology we test our theories against particular cases and if we find the implications of the theory unacceptable in a particular case, we need to revise the theory in a way that avoids those implications. Although there is no clear

line between constructive theory revision and the introduction of fudge factors, there is an important idea underlying the distinction. The motivation for constructive theory revision is to find the best explanation of particular cases. The motivation for the introduction of a fudge factor is to insulate a theory from refutation by particular cases. If this kind of motivation were transparent, it would always be able to tell the difference by introspecting one's own motivation. But biases of these kinds are not introspectible (Talbott 1995). So it is often easier to detect a fudge factor from the outside. Someone who has no allegiance to any of the competing theories in a domain is often in the best position to distinguish between constructive theory revisions and fudge factors.

In any case, it is now clear that a defense of Mill's experimental society over Plato's beehive society on the basis of higher and lower pleasures is unsuccessful. Nozick's experience machine thought experiment undermines all hedonistic accounts of well-being, including Mill's. If there is a defense of the superiority of Mill's experimental society over Plato's beehive society, it will not be so quick and easy.

Once it is seen that autonomy itself does not guarantee a life with higher net hedonic value, it becomes an interesting empirical question whether the experimental society has higher levels of well-being than a beehive society. To his credit, in chapter 3 of *On Liberty*, after the initial argument just discussed, Mill went on to make a surprising and, in retrospect, revolutionary empirical argument in favor of the experimental society. The main idea of the argument is that there is no *a priori* method for determining what human well-being is or how to attain it. The nature of well-being must be discovered by individual experiments in living. But the experiments have value only if individuals' judgments of their own good are generally reliable. And this requires that certain autonomy rights be guaranteed.

In the first volume, I developed this argument more fully (2005, chap. 6). It is only in this volume that I have the opportunity to plumb its depths, which I do in chapter 7. However, there are a number of arguments in the literature that, if successful, would show that a consequentialist account of the kind that I propose here is untenable. In the next two sections I consider and reply to some of the most influential of those arguments.

Objections to Basing Morality or Justice on Well-Being

In this section, I consider the three main reasons that have been offered to favor a theory based on resources or on capabilities over a theory based on well-being. The first is the problem of making interpersonal comparisons of well-being. Without a definition of well-being, I cannot offer a formula for such comparisons. Fortunately, it is possible to resolve many important real-world cases without any simple formula or precise method for making interpersonal comparisons. I provide many examples in this and subsequent chapters.

The second reason that has been offered to favor a theory based on primary goods or on capabilities is the problem of adaptive preferences. Someone who has very limited opportunities may accommodate herself to her situation by diminished expectations, and thus may be relatively content with her lot. This shows that well-being should not be identified with preference satisfaction. Because I do not identify well-being with preference satisfaction, the problem of adaptive preferences does not favor a theory based on resources or on capabilities over my account based on well-being.

The third problem is the problem of encouraging expensive tastes. I can illustrate it with an example: When they were young, both Joe and Oscar enjoyed tofu. Joe still enjoys tofu, but Oscar has voluntarily cultivated a taste for caviar. Caviar is much more expensive than tofu, but Oscar no longer enjoys eating anything else. It would seem that a well-being-based account of justice would favor providing more resources to Oscar than to Joe, because Oscar needs more resources to be able to buy the caviar that will make him as happy as Joe is when he is eating tofu.

Accounts based on resources or capabilities avoid any such implication because they evaluate Joe and Oscar on the basis of factors (e.g., income or opportunities) that are independent of how happy they are with what they have. Thus it seems that well-being-based accounts will favor what most people would regard as an injustice (e.g., providing more resources to Oscar and fewer to Joe) in cases of this kind.

But not all well-being-based accounts have this problem. In fact, this problem provides one more way of appreciating the advantages of indirect over direct consequentialism. The example of Oscar and Joe illustrates why a social practice consequentialist would not favor the social practice of trying to equalize happiness or well-being. The foreseeable effects of such a practice would be to reward individuals for acquiring expensive tastes, and thus to *lower* the overall level of well-being within the society.

This phenomenon, which will be important in coming chapters, is called *insurance effect* (the opposite of what is called *moral hazard*). The insurance effect is the phenomenon that providing insurance against a bad outcome tends to increase the frequency of the bad outcome, because insurance reduces people's motivation to avoid the bad outcome. Sometimes a practice of providing insurance produces good effects that outweigh the bad consequences of the insurance effect. So providing some kinds of insurance can be justified by the social practice consequentialist. But the social practice of equalizing well-being would have no such countervailing good effects. A practice of rewarding those who cultivate expensive tastes with more resources would tend to encourage people to cultivate expensive tastes. In a situation of fixed, scarce resources, any increase in the number of those who cultivate expensive tastes will decrease the extent to which everyone's tastes can be satisfied. Thus, no such practice would be favored by any reasonable form of indirect consequentialism.

In a discussion of the problem of encouraging expensive tastes, R. Dworkin suggests that the kind of solution to the problem that I have presented here is the wrong kind of solution to the problem, because it involves balancing efficiency considerations against equality considerations (2000, 54–55). He believes that an adequate solution should explain why encouraging expensive tastes offends against equality considerations alone.

Here Dworkin fails to acknowledge the depth of a social practice consequentialist account. The social practice consequentialist cannot be expected to explain the *truth* of nonconsequentialist claims. Rather, the social practice consequentialist will try to explain why it is *morally appropriate* to believe the nonconsequentialist claims. It is true that we intuitively judge that it is unfair that Oscar receive more resources than Joe. The indirect consequentialist claims to be able to explain why it is morally appropriate that we make that intuitive judgment. Though the judgment to be explained is not a judgment of efficiency, it could well be that factors related to efficiency help to explain why it is morally appropriate for us to make it.

Of course, efficiency is not the whole story. On the indirect consequentialist account I favor, it is equitably distributed well-being that social practices should promote. But the fact that a practice *reduces* everyone's level of well-being is enough to almost guarantee that it is not a good way of equitably promoting well-being.

A Circularity Problem

R. Dworkin has another argument against well-being-based accounts of justice that would also apply to my account. The problem is that the most promising account of well-being for a life, which he calls the *model of challenge*, makes well-being itself dependent on justice. Roughly, the idea is that there is no adequate measure of the goodness of a life that does not include considerations of justice or fairness, as what Dworkin refers to as a "soft parameter" (2000, 266, 278–279). Call this the *circularity problem for consequentialism*.[5]

To avoid this circularity problem, it is necessary to make a distinction between two ways of talking about well-being. I call them the *broad* and *narrow* conceptions. I believe that Dworkin is correct that, *broadly understood*, the goodness of one's life depends on considerations of justice. It is possible to imagine two nearly identical lives of artistic creation. In both cases, the artist is supported by patrons. In one scenario, the patrons are slave owners, whose profits are derived from the exertions of slaves who live lives of severe deprivation. In the other scenario, the patrons are citizens of a just society, in which everyone has the opportunity to engage in well-compensated, productive work. Suppose that the patrons in the just society are not especially wealthy. They just choose to spend their discretionary income supporting the arts. I think Dworkin is correct that the artist in the just society has a better

life than the artist in the slave society. In this evaluation, the goodness of a life is understood broadly.

If my indirect consequentialist account of moral improvement relied on a broad conception of well-being, it would be hopelessly circular. To avoid such circularity, I must employ a morally neutral conception of well-being. This would be a *narrow* conception of well-being. On reflection, it seems clear that there is some appropriately narrow conception of well-being to play this role, even if no one has given a satisfactory definition or other account of it.

Although we have no complete account of human well-being in the narrow sense, we do know a lot about it. For example, even if the life of a healthy villain is bad in the broad sense, we have no difficulty in understanding why health itself is good in the narrow sense, even for villains. In fact, we make use of our understanding of well-being in the narrow sense in many different contexts—for example, in the design of criminal sanctions. It may well be true that, in the broad sense, the life of a murderer is not a good one. No one in their right mind would base criminal sanctions on well-being in this broad sense, because no one would seriously suggest that we regard murder as its own punishment. Those who evaluate their well-being in the broad sense would not commit a murder even if there were no punishment for it. Sanctions for murder are aimed at those who evaluate their well-being in the narrow rather than the broad sense, to motivate them not to commit murder. Thus, in designing a sanction for murder, we look for penalties (e.g., incarceration) that are bad in the narrow sense.

Political philosophy would be impossible if there were genuine doubt about whether, generally speaking, it is not better for normal human beings to avoid an early death or a life of torture or a life of incarceration. Nor are all uncontroversial judgments of well-being limited to cases as extreme as these. Generally speaking, it is better for most people to spend less rather than more time sick or stuck in rush hour traffic or cleaning up after a pet that has not been housebroken. In this book, for the most part, I limit my discussion to relatively uncontroversial examples involving judgments about well-being in the narrow sense. So the circularity problem for well-being is not a problem for my account.

I conclude here that a consequentialist account cannot be eliminated at the outset. In the remainder of this chapter I further explain the important concepts in the main principle.

Life Prospects

The main principle uses life prospects as a proxy for well-being: *Life* prospects, because social practices must be evaluated in terms of their contribution to the well-being of an entire life—thus, by reference to their contribution to what I refer to as *lifetime utility*. It is important that social practices be

evaluated by their effects on lifetime utility, because a human life can have value as a whole that is more than the sum of the value of its parts. This is because, for human beings, well-being is in part a function of forming and pursuing a plan for one's life.[6]

Because any individual's lifetime utility is a product not only of social arrangements but also of various natural and other contingencies, social arrangements cannot guarantee any actual level of lifetime utility. For example, changes in social practices can increase life expectancy, but they cannot guarantee everyone a long life. So changes in social practices must be evaluated in a way that allows for incorporating risk and uncertainty. By life *prospects*, I mean to refer to a measure of lifetime utility that incorporates risk and uncertainty.

The usual measure of life prospects employed by consequentialists is the expected utility function, but there is in the literature reasonable disagreement about how to define expected utility and even whether it can be reasonable not to maximize expected utility.[7] Because I see no decisive consideration in favor of any one of these various alternatives, I cannot assume that the main principle defines *life prospects* in terms of expected lifetime utility. Even in the absence of a specific formula for life prospects, we can at least identify the relevant variables that determine them. A person's life prospects at a particular time are determined by a probability distribution over lifetime utilities.[8]

There are two more things to be said about the concept of life prospects as employed by the main principle. First, unlike most consequentialist principles of distributive justice (including both Mill's and Rawls's), the main principle is not an averaging principle. Averaging principles evaluate a practice by averaging the life prospects of many different individuals (in the case of average utilitarianism, by averaging the life prospects of all individuals). In contrast, the main principle evaluates a social practice by its effects on the life prospects of each individual.

Second, unlike most consequentialist principles, the main principle does not limit its evaluation of life prospects to a particular *cutoff date* (e.g., birth or adulthood). It evaluates social practices by their effects on the life prospects of everyone at every stage of life.

Consider an example. Suppose Donald had very good life prospects at age 25 when he entered into a contract to work in a dangerous occupation. Donald knew that the occupation was dangerous. He chose it because of the premium pay for the element of danger. Suppose that the premium pay was fair compensation for the risk. Unfortunately, at age 30 Donald was injured at work and disabled for the rest of his life. This would be an example of what Dworkin calls *option luck*, because Donald knew of the risk and accepted it (2000, 73). Many nonconsequentialist theories would say that if the background institutions are just, then bad option luck, as in Donald's case, raises no problem of justice.[9] Most consequentialist theories would give the same result, typically because they would average Donald's life prospects

with the life prospects of others or because they would evaluate Donald's life prospects at a particular cutoff date (e.g., at birth or at age 25 when his life prospects were very good).

The main principle does not yield this result. The main principle does not average life prospects across individuals and does not use a cutoff date for measuring life prospects. Donald's life prospects at age 25, before his injury, may have been very high, but the main principle will also take into consideration his life prospects at age 30, after the permanently disabling injury. Under the main principle, the distribution of lifetime utilities at age 30, and indeed, at every age, matters.

Because I have no definition for life prospects, in my discussion of the main principle, I will have to rely on examples in which it seems plausible that any reasonable measure of life prospects would yield the same ranking. Consider a simple example. Suppose there is an immunization against a common fatal disease that could significantly reduce everyone's chances of contracting the disease with no adverse side effects except the momentary pain of the injection. Suppose however that, in order for the vaccine to be effective, it is necessary that the entire population be immunized and that it is very improbable that everyone would voluntarily seek the immunization if there were no punishment for not being immunized. If there were a system of forced immunization that did not have other significant negative effects, when substantively evaluated under the main principle, the system of forced immunization would generally rank above a system of voluntary immunization, because, in most circumstances, on any reasonable ranking of life prospects, everyone's life prospects would be higher at every stage of life under the system of forced immunization than under a system of voluntary immunization.[10]

Promoting Life Prospects

The main principle evaluates social practices in part by the extent to which they *promote* life prospects. The standard for improvement is the status quo system of social practices. In addition, the potential improvement must be ranked higher than any of the relevant alternatives. I have no general account of when an alternative is relevant. I simply rely on our ability to pick out the relevant alternatives in particular cases.

Consider a variation on the example of the forced immunization. Suppose there is a fatal disease for which there is no vaccine, but the fatality rate can be greatly reduced by a system of forced quarantine of those who contract the disease. It is quite plausible that, in comparing the forced quarantine system with a system of voluntary quarantine, the system of forced quarantine would increase everyone's life prospects at birth (before anyone had contracted the disease). However, the main principle applies to people's life prospects at every stage of their lives. Those who contract the disease will have their life

prospects diminished by the forced quarantine over what they would be without a forced quarantine. Because I take it to be clear that in many cases the forced quarantine should be favored over the system of voluntary quarantine (on the assumption that those who were quarantined were well treated), this example would seem to favor a simpler consequentialist principle, either an averaging principle or one with a cutoff date at birth. Almost any such principle would directly yield the result that the system of forced quarantine was superior.

The main principle does not give the correct result so directly, because it requires some way of trading off advantages to some against disadvantages to others. There would be no reason to accept the more complex main principle if the simpler principles were adequate for all cases. Unfortunately, as I illustrate shortly, the simpler principles are not at all adequate. So, in making evaluations under the main principle, there is no way to avoid tradeoffs. When there are tradeoffs to be made, equity becomes an important consideration in the application of the main principle.

Equitably Promoting Life Prospects

The main principle makes the evaluation of a social practice dependent on the extent to which it *equitably* promotes the life prospects of cooperators. Probably the most imprecise notion in the main principle is the notion of equity. What is required for a practice to *equitably* promote the life prospects of cooperators? I have no precise answer to this question, because I have no formula for equity. Again, I must rely on our ability to make reliable judgments in at least some particular cases. There is a heuristic that is useful for helping us to make such judgments, an *expanded original position heuristic*.

The Expanded Original Position Heuristic

In *Theory of Justice*, Rawls introduced a device that he hoped could be used to derive the principles of distributive justice, the *original position*.[11] Scanlon (1982) questioned whether the construction, as Rawls defined it, could ever be used to derive principles of justice, but he allowed that something like it might be useful as a heuristic. I propose to take up Scanlon's suggestion and use a relative of Rawls's construction as a heuristic to help us to decide when one social practice does a better job of equitably promoting life prospects than another.

In Rawls's original position, the parties are imagined to go behind a veil of ignorance that deprives them of all particular, individuating information about themselves. They do not know their social position, their comprehensive view (including their religion), their race, nationality, family, the generation to

which they belong (although they are assumed to be contemporaries) (1993, 273–274) or any other information that would enable them to distinguish themselves from any of the other parties in the original position (1971, 137; 1993, 25). They do, however, possess all general information (1971, 142).

There are four main differences between my heuristic and Rawls's construction. First, unlike Rawls, I do not assume that the parties are merely rational in Rawls's sense (1971, 142). I assume that they are trying to be fair. The original position simply introduces an impartiality constraint that helps them to figure out what fairness requires.

Second, because I use the heuristic to help decide real-world cases, I do not limit its use to the basic structure of society. I use it to evaluate changes in social practices, both moral and legal.

Third, I do not limit the use of the heuristic to the kinds of ideal situations that Rawls discusses—for example, the ideal situation in which everyone is willing to cooperate on fair terms of social cooperation or in which it is assumed that there is an overlapping consensus on the principles that would be selected in the original position. Suppose that it was general knowledge that 20% of the members of a society were not willing to cooperate on fair terms of social cooperation. It might be disastrous to adopt laws that everyone would agree to in Rawls's original position, if it is known that 20% of the society would not comply with those laws. In my expanded original position, the parties are those who would be willing to comply with the practices to be evaluated (the *cooperators*). It is the cooperators whose life prospects are to be equitably promoted by the relevant practices. For example, no effective prohibition on murder could ever be justified, if it had to promote the life prospects of everyone, including murderers.

Fourth, unlike Rawls, I do not exclude those with special health care needs or physical or psychological impairments, whether temporary or permanent, from consideration (1971, 510; 1993, 20, 21, 25, 272 n. 10). I include them with the other nonresponsible noncompliers as cooperators.

The expansion of the parties in the original position to include those with special health care needs raises a theoretical problem, because, clearly, someone with a severe brain impairment could not carry out the reasoning involved in this modified original position thought experiment, and thus it would seem that from the mere fact that they could carry out the reasoning, the parties could know that they did not have severe brain impairments.[12] Although this would be a problem for a Rawlsian construction, it is not a problem for my version of the original position understood as a heuristic. A heuristic is not a decision procedure. It may be useful in some cases and not in others. I think that it is possible to consider those with severe brain impairments in the original position by asking yourself what agreement you would think was fair, on the understanding that, after the veil of ignorance is lifted, you might develop a severe brain impairment.

I refer to the original position heuristic as modified above as the *expanded original position (EOP)*. Consider how the EOP would apply to the forced

quarantine example discussed above. The expanded original position heuristic makes it easy to see that the tradeoff involved in the example—the advantage of a higher probability of avoiding contracting the disease and living a normal life versus the disadvantage of dying under quarantine in the much less likely event that one contracts the disease—is easily evaluated. The forced quarantine alternative is clearly superior, even though some people (those who contract the disease) would have higher life prospects under the alternative of no forced quarantine. In this sort of case, it seems to me uncontroversial that there would be general if not unanimous consent in the expanded original position. Because consent in the expanded original position is based entirely on a comparison of life prospects under the two alternatives, it is a useful test for endorsement by the main principle.

There is an obvious advantage to including those with special health care needs in the original position: It simplifies the theory of moral improvement if the theory can apply to everyone in a given society, rather than excluding those with special health care needs, as Rawls's theory does. But there seems to be a weightier consideration on the other side. Both Rawls's theory and mine are concerned with reciprocity grounded entitlements (1993, 16–18). But how could those with special health care needs have reciprocity grounded entitlements if they are unable reciprocate cooperation? I answer this question in chapter 11.

Rawls had an additional reason to exclude those with special health care needs from his original position. Because his maximin expectation principle required maximizing the expectation of the least advantaged group, if those with severe mental handicaps were included in the least advantaged group, Rawls's theory could end up requiring that most of society's resources be spent improving their situation, an extremely implausible result. This shows that there is no plausible way to include those with special health care needs within the scope of Rawls's theory. I return to this topic shortly.

Keep in mind that, as I use it, the EOP thought experiment is only a useful heuristic, and then only when applied in good faith. Suppose an antebellum slave owner claims that being a slave is better than being a free man: "Being a slave has lots of advantages. All your needs are taken care of." We suspect that the justification is self-serving. So we can ask ourselves this: Would the slave owner accept his own justification in the expanded original position, in which he would not know if he was a slave or a slave owner? It will generally not help to ask the slave owner this question. If he is acting in bad faith, he will almost surely say yes. He might add, "If I were a slave I would be grateful that I had someone to take care of all my needs."

However, even if the expanded original position thought experiment does not decide all disagreements, it does provide a way of detecting bad faith. Even if a slave owner tells us that he would accept slavery in the expanded original position, it is hard to believe that he is right. It is simply not plausible that he would accept the possibility of being put in a position in which he and his family members could be bought and sold and in which he was

completely at the will of another, with no rights at all. If the slave owner insists that he would accept it, it is very likely that the slave owner's justification is a self-serving and thus bad faith justification.

I will have much more to say about the EOP. However, it is important to note that parties in the EOP are *not* assumed to make decisions on the basis of overall expected utility or any other measure of aggregate measure of utility and they are *not* assumed to limit their comparisons of life prospects to a single cutoff date. They are expected to make decisions based on individuals' life prospects at every stage of life.

This feature makes the main principle more complicated than the comparatively simpler principles of Mill (utilitarianism) or Rawls (maximin). To appreciate the need for the more complex principle, it is useful to discuss the shortcomings of Mill's and Rawls's simpler principles.

Problems for the Utilitarian Formula and for Actual-World Defenses of It

Mill had a consequentialist formula for the justice of social practices—that they maximize overall utility. The early Rawls argued persuasively that Mill's formula was inadequate because of its distributional blind spot (1971, 3–4). Utilitarianism's distributional blind spot is most easily appreciated with hypothetical examples. Even if the most efficient organization of society would require a small minority to make almost all of the sacrifices and everyone else to receive almost all of the benefits, the fact that such an arrangement maximized overall utility would not make it just. If there were an alternative with a much more equal sharing of the sacrifices and the benefits, it might be favored by justice, even if overall utility were not as high.

Utilitarians sometimes object to such hypothetical examples, on the grounds that facts about human beings rule them out. When applied to human societies, it is argued, utilitarianism gives the correct result, because, for example, as a consequence of the diminishing marginal returns to utility from money and other resources, it would not favor practices that produced large disparities in life prospects. I need a name for this sort of tactic for defending a philosophical theory. I call it an *actual-world narrowing*. An actual-world narrowing of a theory does not attempt to defend the theory's implications in hypothetical cases involving possible worlds different from ours; it simply tries to persuade us that the theory gives the right results in cases in the actual world and in possible worlds very similar to the actual world.

What is wrong with an actual-world narrowing of a theory? On the largely bottom-up model of reasoning that I employ, moral theories get their support from their role in explaining actual and hypothetical cases. An actual-world narrowing of a theory implicitly acknowledges that there are hypothetical cases that it does not explain. The defense is supposed to reassure us that

none of those hypothetical cases occurs in our world, so we need not worry that the theory will lead us astray in our world.

Actual-world narrowings of an explanatory moral theory should raise our suspicions. They should make us wonder whether it is just a coincidence that the theory gives the right results in the actual world and why the defender of the theory is so confident that the assumptions on which the actual-world narrowing depends are true of the actual world. An actual-world narrowing is a reason to doubt a theory, not to embrace it.

There is another problem with actual-world narrowings of a moral theory. If a theory has true implications only for actual-world cases, the theory does not actually *explain* even those cases. For example, even if utilitarianism gave the right result in all actual-world cases, its failure in other hypothetical cases would show that utilitarianism is *not* the explanation of the actual-world cases that it gets right. An explanation of the actual-world cases that utilitarianism gets right would have to include an explanation of why utilitarianism gives the right results in those cases (but not others) and, thus, an explanation of why utilitarianism approximates the true theory in actual-world cases. So we would need some other moral theory to explain why utilitarianism is so close to the truth when applied to actual-world examples.

An actual-world narrowing is one example of a broader category of defenses of a philosophical theory, which I refer to as *subclass defenses*. My defense of consequentialism in this book is an example of a subclass defense of consequentialism in moral philosophy. Mill's project was to give a consequentialist account of all morality. Although I have sympathies with Mill's larger consequentialist project, in this book I propose to explain why, even if all of morality is not consequentialist, all or most improvements in primary ground level moral thought can be given a consequentialist explanation. Subclass defenses can be theoretically illuminating if they include a theoretical explanation of why the relevant theory only applies to the narrower subclass class rather than to the larger one. For example, it is at least plausible that there is a difference between primary and secondary moral norms (e.g., norms of distributive justice and norms of corrective justice), and thus that there might be a consequentialist explanation of the former even if there is no consequentialist explanation of the latter.

But there is no reason to think that there is an important theoretical difference between possible worlds very much like the actual world and other possible worlds less like the actual world, so a subclass defense of a moral theory that limits the theory to worlds very much like the actual world should raise doubts about the adequacy of the theory, just as ad hoc narrowing of a scientific theory to avoid unwelcome experimental results should raise doubts about a scientific theory. One of the main problems with actual-world narrowings of philosophical theories is that the distinction between the actual-world and other nonactual possible worlds is almost never theoretically significant, so actual-world defenses almost never offer a theoretically interesting explanation

of why the relevant theory would apply only to the actual world (and other possible worlds very similar to it) and not to other possible worlds more different from it.

It is important not to misunderstand my criticism of actual-world narrowing of a theory. It is not a *defect* of a theory that it explains examples in the actual world. It is a virtue. The defect would be that it applies *only* to the actual world (and worlds like the actual world) and does *not* explain examples in other possible worlds that differ from the actual world.

Suppose, for example, that utilitarianism did apply to all the interesting examples in this world, but that it did not apply to other possible worlds. By exploring its implications for those other possible worlds, we might discover that the reason it applied to all examples in our world was that all the social practices in the actual world that maximized overall utility also distributed it equitably. As moral philosophers, we don't just want a theory that gives us the right results for some examples but not others. We want to understand why it gives the right results in the cases in which it does so and why it does not give the right results in the cases in which it does not. In addition, we should be concerned that the assumptions about this world that are necessary for the theory to apply to it may be motivated by the desire to save the theory and, thus, may not even be true of this world.

Problems with Rawls's Maximin Expectation Principle

On my interpretation, though the early metaphysical Rawls rejected the utilitarian formula, he did not reject the consequentialist project for distributive justice. The early Rawls was a social practice consequentialist about distributive justice, because his general conception of justice, from which he claimed to be able to derive the special conception, was consequentialist. His general conception had a single principle, which I refer to as the *maximin expectation principle*—roughly to maximize the expectations for primary goods of the least advantaged group (1971, 303).[13]

There are two things to be said for the maximin expectation principle. First, it pays attention to the distribution of the good. In questions of distributive justice, distribution matters. Second, in the evaluation of a social practice, it places special emphasis on the good of the least-advantaged group. There is reason to expect a theory of distributive justice to give some sort of priority to those whose shares in the benefits of cooperation are lower than others. However, it must have been clear to Rawls, even at the time he proposed it, that his maximin expectation principle could not be the general consequentialist formula for distributive justice, for even at that time he was aware of some powerful objections to it. I begin with the objections he was aware of and then add one that he does not seem to have been aware of.

First, it was clear to Rawls that the maximin expectation principle cannot apply to people with severe physical or mental disabilities. It is easy to imagine

hypothetical cases in which the maximin expectation principle would imply that people with such disabilities were entitled to practically all of society's resources, because their life prospects are so much below most other people's and it can be very expensive to produce even marginal improvements in their life prospects. The early Rawls addressed this potential objection by assuming that "everyone has physical needs and psychological capacities within the normal range, so that the problems of special health care and of how to treat the mentally defective do not arise" ([1975], 259). Limiting his theory in this way was Rawls's attempt to give a subclass defense of his theory. Rawls had a good reason for this limitation. He wanted a theory of distributive justice for a scheme of social cooperation to which everyone was assumed to contribute, because the main issue of distributive justice for Rawls was how to fairly distribute the jointly produced social product (15).

If Rawls's maximin expectation principle had at least applied generally to those without special health care needs, it would be an important principle of distributive justice, even if not a comprehensive one. But even when he wrote *Theory of Justice*, Rawls was aware of the need for an additional subclass defense. The problem for Rawls's principle can best be illustrated by a hypothetical example. Rawls's principle requires maximizing the expectation (of primary goods) for the least advantaged group, regardless of the potential losses to other groups. But even if we set aside special health care needs, we can imagine a situation in which maximizing the expectation of the least advantaged group requires large reductions in the expectations of other groups for only marginal improvements in the expectation of the least advantaged group. It is very implausible that justice would require large reductions in majority expectations merely to marginally improve the expectations of a minority, even if the minority is the group with the lowest expectations. This argument becomes stronger if we suppose that the expectation of the least advantaged group is for a comfortable life free of hardship.

Rawls anticipated this objection and responded to it. How did he respond? Rawls simply claimed that "the possibilities which the objection envisages cannot arise in real cases" (1971, 158). So Rawls gave an actual-world narrowing of the maximin expectation principle.[14] But, for the reasons discussed above, we should be skeptical of this kind of defense of his theory. Even if it is true that the facts in the actual world are such that the maximin expectation principle gives the correct result in actual-world cases, we would still need an explanation of why it gives the correct result in those cases. That explanation would presumably be an explanation of why Rawls's principle gives the same results in the actual world as the correct principle of distributive justice. So by giving an actual-world narrowing of his principle, Rawls was implicitly conceding that it is not the correct principle. There is more work to be done.

There is a third objection to Rawls's maximin expectation principle that I don't believe Rawls was aware of. It seems to me to be decisive against his principle, because it shows that Rawls's own principle has the same sort of distributional blind spot that was fatal to utilitarianism.

To appreciate the distributional blind spot of the maximin expectation principle, note that it evaluates the status of each relevant group by their expectation for primary goods, which is an average over the entire group. For Rawls, there are lots of primary goods. I can illustrate the problem for Rawls's principle by using money, which is one of Rawls's primary goods, as a proxy for all of them. The problem with Rawls's principle is that it seems to be a single-time-slice principle. Rawls says that his principles are intended to regulate the kinds of inequalities that favor some starting places over others and are present at birth (1971, 96). So I assume that the expectations are calculated at birth. The argument goes through no matter when the expectations are calculated. I choose birth as the cutoff, because it is the most plausible alternative.[15]

To see why this is a problem, imagine a society in which each person's equal basic rights are guaranteed. The issue is the justice of the society's economic system. Suppose that the society contains only two classes, the least advantaged group (LAG) and the most advantaged group (MAG). In Rawls's theory, each group is represented by one number, the expectation (average) at birth of the primary goods enjoyed by the members of the group over the course of their lives.

So we imagine that there are two groups with different expectations of primary goods at birth. Suppose the expectation of the LAG is 50 and the expectation of the MAG is 100. Of course, expectations are averages, and not everyone in each group will receive the average (Rawls 2001, 173). Some will receive more than the average and some less. The variance from the average could be very great in the MAG and very small in the LAG. For example, if the members of the MAG are those with entrepreneurial characteristics that make them risk takers, perhaps some of them will become billionaires and others will go bankrupt and live in poverty. If those in the LAG lack the entrepreneurial characteristics that would make them risk takers, they might all live a comfortable life free of hardship, with no one becoming very wealthy and no one sinking into poverty. From this example, it is easy to see that some members of the MAG might actually receive less in primary goods over the course of their lives than *any* member of the LAG, as they would if some members of the MAG fell into poverty but no member of the LAG did.

Suppose that the society is considering whether or not to spend tax money to provide antipoverty programs for the unfortunate members of the MAG. It is easy to see that they might not be justified by the maximin expectation principle. To provide antipoverty programs would raise the expectation of the MAG, for example, from 100 to 102. But the maximin expectation principle would oppose any such expenditure if there were an alternative use of the taxes that would raise the expectation of the LAG, for example, from 50 to 52. It would almost certainly be possible to implement some program that would increase the expectation of the LAG, so maximin would not permit funding any antipoverty programs, even if those members of the MAG who went bankrupt would receive a level of primary goods over their lifetime that was much less than 50.

The example is overly simple to illustrate the structural problem with the maximin expectation principle, which is that it is an averaging principle. Although it would be possible for the advocate of the maximin expectation principle to object to the details of my particular example, it is harder to see how to reply to the structural problem that it illustrates. For example, someone defending the maximin expectation principle might reply that if a member of the MAG went bankrupt, he or she could obtain help from family members who would also presumably be members of the MAG. But this might not be true. It might be that all businesses were family owned and family run, so that when a business went bankrupt, it bankrupted the entire family. Or it might be that the members of the MAG came from extremely competitive families whose members would not help each other. Of course, it will still be possible to move to an actual-world narrowing by denying that any such case should arise in the actual world. From a theoretical point of view, such a defense is little more than an acknowledgment of the need for a better principle.

Once this structural problem in the maximin expectation principle is recognized, it is apparent that that principle has the same kind of distributional blind spot as the utilitarian principle.[16] One way of trying to fix the problem with the maximin expectation principle is to revise it so that the maximin rule applies not to expectations, but to actual outcomes. The revised principle would require maximizing the actual level of primary goods of the least well off person. Call this the *maximin outcome principle* to contrast it with Rawls's maximin expectation principle. The maximin outcome principle is much less plausible than Rawls's principle, because it would rank an egalitarian society in which everyone achieved the same modest level of primary goods (e.g., 50) above a society in which everyone had an expectation of a much higher level (e.g., 1,000), if even one member of the latter society actually fell below 50. This is much too extreme to be an adequate principle for distributive justice. There is another problem, which I will illustrate shortly.

Another suggestion for solving the distributional blindspot of the maximin expectation principle is not to apply it at a single cutoff date, but rather to apply it at every instant throughout a person's life. Call this principle *continuous maximin expectation*. Continuous maximin expectation would require that society maximize the expectations of those in the least advantaged group over everyone's lifetime. So if at a late age someone fell into the least advantaged group, their expectations would be maximized. It should be obvious that continuous maximin expectation is unacceptable, for the same reasons that maximin outcome is. Setting aside issues of special health care needs or disability, who would compose the least advantaged group over the course of their lifetime? The group would probably include those who are regularly fired by their employers, have multiple divorces, and multiple bankruptcies. I believe a strong case can be made for providing job training opportunities and for marriage counseling and financial

education, but that is very different from maximizing the expectations of those who are regularly fired, divorced, and bankrupted. Equity surely requires some kind of safety net for the members of this group. I see no reason that it should require maximizing their level of expectation. Though there are some unfortunate exceptions, most people who are often fired, divorced, and bankrupted are irresponsible and inconsiderate. No acceptable theory will require raising their expectations so high as to encourage irresponsibility and inconsiderateness. This seems to me to be a decisive consideration against the continuous maximin expectation principle and against the maximin outcome principle, also.

There is one more maximin principle that can be considered briefly. If Rawls's maximin expectation principle were applied to the entire population, including those with special health care needs, it would probably require that large amounts of resources be devoted to the care of those with severe brain impairments, even if benefits were only marginal. However, it is much more plausible to think that justice requires that a legal system minimize the number of those who suffer severe losses (e.g., severe brain impairments), not necessarily maximize the expectations of those who suffer such losses. Call this the *maximin loss principle*. The problem with the maximin loss principle is that it would justify prohibiting many if not most risky activities, including most sports. It is very implausible that distributive justice requires prohibiting most sports. A separate question, which I take up in chapter 11, is whether distributive justice requires any aid to those who are injured in such activities. There I argue that the answer is yes.

The discussion so far suggests that no simple principle of justice can be adequate. On the one side are principles defined in terms of expectations on a given cutoff date. Such principles all have a distributional blind spot. On the other side are principles that pay attention to actual outcomes or to expectations throughout an entire life. If those principles are defined lexically (e.g., maximin outcome or maximin loss), they are too strict. They do not allow for activities in which risk is an essential part of the activity.

The main principle occupies a place between these two extremes. Because the main principle does not average life prospects and does not limit the consideration of life prospects to a single cutoff date, it does not have a distributional blind spot. However, it does not employ any kind of lexical ranking, because, in some cases, it permits practices that produce very bad outcomes, even if there are alternatives that would reduce the incidence of those very bad outcomes.

Sufficientarian, Egalitarian, or Prioritarian Equity?

I conclude that no version of the maximin principle will be adequate for a consequentialist account of distributive justice, or for an account of moral improvement that covers beliefs about justice. There are in the literature three

other kinds of formula that seem promising: sufficientarian accounts, egalitarian accounts, and prioritarian accounts. I discuss them individually.

A sufficientarian account replaces Rawls's imperative to maximize the expectation level of the least advantaged group with a less onerous imperative to raise each group above some morally urgent baseline, typically defined in terms of needs (e.g., Frankfurt 2000). All such accounts face the problem of explaining how there could be any such baseline. The pressures come from two directions. On the one hand, the baseline has to be so important that it would make sense to think that raising everyone above it should be an absolute social priority. For this to be plausible, the baseline would have to be fairly low (e.g., subsistence or near subsistence). On the other hand, it is hard to believe that society has no duties of justice to those who are above such a low baseline—for example, no duties to create opportunities for them to rise above it.

The main principle is not sufficientarian. The main principle's standards of equity continue to apply even after everyone's life prospects are above any reasonable sufficientarian baseline. Indeed, the main principle always gives more weight to opportunities for the less well off than to opportunities for the more well off, no matter what the absolute level.

Egalitarian theories are those that assign some positive value to equality of well-being per se (e.g. Temkin 1993). It might seem that such a theory would be admirably suited to providing a theory of equity, because there does seem to be some moral appeal to the idea of equality of life prospects. The problem is that because some people's incapacities greatly limit their life prospects, a theory that gives positive value to equality will have some tendency to favor *reductions* in the well-being of the more well off, even if those reductions don't make any improvement in the level of well-being of the least well off, because even if they don't improve the level of the least well off at all, they still reduce inequality. Parfit (1997) calls this the *leveling down objection*.[17] This seems to me to be a decisive objection to egalitarian theories as theories of equity.

The main principle would never rank a system of social practices above an alternative that was Pareto superior to it. Any improvements in life prospects are always a net gain. However, the main principle can endorse a change that reduces some people's life prospects. The elimination of a slavery practice may reduce the life prospects of some of the slave owners. The main principle could endorse it if the nonslaveholding practices more equitably promoted everyone's life prospects.

Also, the main principle weights improvements in the life prospects of the less well off more heavily than improvements in the life prospects of the more well off. This makes the main principle *prioritarian* (Nagel 1991, 69; Parfit 1997, 213). However, because it is an indirect consequentialist principle, it is prioritarian in a more nuanced way than usual, as illustrated by this example: Imagine two people, Helen and Ray, who are both paralyzed from the waist down. Helen's condition was congenital. The system of social practices had

no effect on the incidence of Helen's kind of paralysis. Ray was paralyzed in an auto accident in a car with defective brakes.

On most prioritarian views, there would be no reason to single out Ray over Helen for special compensation. In fact, given that Helen's condition was congenital, a prioritarian social welfare function might well rank her prospects *lower* than Ray's and thus *favor* granting Helen more compensation than Ray for disability. It is somewhat surprising that the main principle can endorse discrimination in the other direction. It can endorse a system of disability insurance that provides Helen with a level of compensation *less* than the level of compensation that Ray would be entitled to recover from the manufacturer of the car with defective brakes. Why does the main principle pay attention to the source of paralysis in this way? The guiding idea is that Ray's paralysis is one of the costs of the social practice of driving cars. Because the main principle evaluates life prospects at every stage of life, it can keep track of those whose life prospects are diminished by a social practice and can favor internalizing the costs of a social practice, so that they are paid by those who benefit from the practice. This is a topic that I return to in chapters 9 and 11.

More on Special Health Care Needs

An important difference between my theory and Rawls's is that my theory is intended to apply to those with special health care needs. First, let me say why it is a mistake to think that a theory of distributive justice such as Rawls's could ignore the life prospects of all people with severe and permanent disabilities. Rawls's theory would evaluate social practices on the basis of whether they maximize the expectation of the least advantaged group, where everyone is assumed to be "normal and fully cooperating members of society over a complete life" (2001, 83) But no theory can be adequate if, like Rawls's, it ignores the life prospects of those who are permanently disabled by their participation in hazardous social practices, such as fire fighting. A theory such as Rawls's would yield an unacceptable result in the following kind of case. There is a high-paying occupation that greatly raises the expectation for primary goods of those who engage in it. However, X% of those who engage in this occupation are seriously disabled. As Rawls applies his maximin expectation principle, this occupation would make the society more just. It would raise the expectation level of the able-bodied, *regardless of how high X was*, because those who were disabled by the occupation would not be considered in the evaluation of it.

Of course, Rawls's principle is meant to apply to the basic institutions of a society, not to individual occupations. But the example shows a structural problem with the principle that would carry over even to applications to the basic institutions of a society.

Examples of this kind persuade me that the life prospects of all the members of a society who are not covered the responsible noncompliance exclusion need to be considered in the main principle's evaluation of social practices, even those who are disabled, and especially those who are disabled as a result of their participation in hazardous practices. When a social practice compensates those who suffer disabling injuries from their participation in it, that raises its ranking under the main principle.

Rawls seems to think that distributive justice should cover only those who have actually participated in the scheme of cooperation over the course of a life, but this seems to me to be a mistake. To modify the previous example, suppose that there is an occupation that carries with it a significant risk of severe disability not for the worker, but for his or her children. On Rawls's account, the children who were born disabled would not even count in the evaluation of the basic structure of the society, because they would never be normal participants in the system of social cooperation.

Of course, Rawls would not deny that we have a duty to assist those children. He would hold only that we have no duty of distributive justice to them. This can't be right. Consider a system of basic institutions, S_1, in which, due to the hazardous chemicals used in production, many children are born severely incapacitated. After providing welfare benefits to the disabled children necessary to satisfy whatever duties the members of S_1 have to them, S_1 is wealthy enough to assure the LAG of able-bodied workers a wage of $20 per hour. Now compare S_1 with S_2. S_2 does not use any hazardous chemicals. It is a less wealthy society, but no children are incapacitated. The maximum wage that S_2 can guarantee to the LAG of able-bodied workers is $18 per hour. Rawls's theory would imply that S_1 is more distributively just than S_2. This is extremely implausible. The incapacitated children are bearing some of the burdens of the social cooperation in S_1. They should be included in the consideration of whether the distribution of benefits and burdens is fair. If we consider them, S_2 is clearly superior to S_1.[18]

Rawls could not include those with special health care needs in his theory of justice, because it would have been obvious that his package of primary goods was not an adequate proxy for well-being for them and it would not be plausible to apply any reasonable kind of maximin principle to them. As I explain in chapter 11, some provision for those with special health care needs and those with severe disabilities would be endorsed by the main principle, but there would be no presumption that their life prospects or any other reasonable proxy for their well-being should be maximized.[19]

The Guiding Idea behind the Main Principle

Rawls's theory of distributive justice is a theory of the fair distribution of the benefits and burdens of cooperation in moral reciprocity relations. The main principle extends this idea to most if not all primary ground-level moral

norms, at least in moral traditions that have crossed the consequentialist threshold.

One reason that the main principle does not lead to counterintuitive results when applied to those with special health care needs or with severe disabilities is that it does not treat the benefits and burdens of a cooperative social practice as something to be distributed on the basis of need. The main principle regards the benefits and burdens as internal to the practice, to be shared fairly by those who cooperate in the practice. The general presumption is that if the practice is worthwhile, all or almost all who cooperate in it should benefit from it in an equitable way. The crucial question then is to determine the extent of those who cooperate. In chapter 3, I included compliers and nonresponsible noncompliers. Here I include all those members of the relevant group who would be expected to be willing to cooperate as compliers, even if circumstances (e.g., congenital disabilities) prevent them from actually cooperating or limit significantly the extent to which they are able to cooperate. Thus, even those who are disabled from birth are covered by the main principle and are included in the expanded original position.[20]

Are There Universal Standards of Equity?

It is sometimes claimed that standards of equity are not universal, but parochial. Ironically, this claim has often been made by Marxists, who claim that capitalist societies generate a capitalist conception of equity. This Marxist claim is undermined by the fact that the appeal of Marxism itself is based at least in part on the claim that workers in capitalist societies are denied their fair share of the benefits they produce by their labor—that is, that the distribution of the costs and benefits of capitalist production is not equitable. One of the most plausible explanations of the worldwide success of Marxism in the twentieth century is that people everywhere shared implicit standards according to which the distribution of benefits and costs between capitalists and workers at the end of the nineteenth century and early in the twentieth century was not equitable.[21] Here again, it is important not to confuse levels. Even if there is no exceptionless ground-level standard of equity, there may be a meta-level principle that explains the appropriateness of judgments in particular cases.

Skyrms (1996) has suggested that there may be an evolutionary explanation for cross-cultural agreement on judgments of equity (and inequity). The evolutionary explanation would not require that there be any truth about matters of equity. There would be only evolutionarily produced agreement. The main problem with Skyrms's proposal is that it implies that there is no standpoint from which to morally evaluate the standards of equity that evolution may have favored. But it seems that some of the standards of equity favored by evolution—for example, patriarchal standards

of gender equity—are not truly equitable.[22] Evolution gave us the capacities we need to be able to make judgments of equity; it did not determine the content of those judgments.

Because I lack a formula for even calculating life prospects, much less a formula for comparing them for distributive equity, it might seem that my theory would be pretty useless. However, it turns out that the existing systems of social practices affect life prospects in such an inequitable fashion that it is easy to find potential changes to the status quo that would be favored by any reasonable measure of the equitable promotion of life prospects. Again, the expanded original position is a useful tool for making such judgments. As I illustrate in the subsequent chapters, the theory gives determinate results in a large number of interesting cases.

The Main Principle Evaluates Social Practices, Not Norms or Principles Per Se

The main principle requires that exceptions to the status quo be evaluated as social practices, both as substantive social practices, on the assumption that they have been implemented, and as social practices of implementation. For simplicity, I continue to focus on the substantive evaluation.

Why does the main principle evaluate social practices and not, for example, evaluate ground-level norms or principles based solely on their content? The answer is that, from the point of view of the main principle, ground-level norms or principles have value only to the extent that they are part of social practices that actually motivate behavior that tends to equitably promote life prospects. Consider a human social practice with a norm of keeping one's promises. Suppose it is suggested that the promise-keeping norm be replaced with the following norm: "Keep your promises, unless by breaking them you would equitably promote life prospects." Surely, one might think, the main principle would endorse a change in the promise-keeping norm to allow such an exception. However, the change would never be endorsed by the main principle, because the effect on the *practice* of promising of allowing people to make such an exception to the promise-keeping norm would almost surely be to greatly reduce the benefits of the practice, which depend in large part on its being reasonable to rely on others to keep their word. Ironically, the practice of permitting people to break their promises when they thought doing so would equitably promote life prospects would itself generally *reduce* life prospects. It is the consequences of the practice itself that determine the main principle's evaluations of it, not the consequences of ideal compliance with the norms embedded in the practice.

There is another reason that the main principle does not evaluate ground-level principles or norms per se. As I explain in chapter 5, the main principle actually favors practices that employ defeasible ground-level norms and

principles over practices that employ categorical ones. If we were to limit ourselves to evaluating ground-level systems of norms or principles, even on the assumption of ideal compliance, we would fail to consider alternatives that would rank higher under the main principle than *any* system of ideal compliance with categorical ground-level norms or principles. Noncategorical norms and principles enable human societies to evolve social practices that, over time, can do a better job of equitably promoting life prospects than any system of categorical norms and principles would. No matter what system of ground-level norms or principles a human society may have, there is always a potential for improvement. This is a surprising result that I explain in chapter 5.

Solution to the Problem of Descriptive Relativity

Because the main principle evaluates social practices rather than norms or principles per se, it avoids a puzzle for many moral theories—what might be called the *problem of descriptive relativity*. Consider a simple example. The main idea of Kantian universalizability accounts of morality is that one be able to universalize the maxim of one's action. For such accounts to be remotely plausible, it makes a big difference how those maxims are stated. Stealing whenever I feel like it is a maxim that most people would not be willing to universalize (when the consequences are taken into account), but what about the gerrymandered maxim that people who are born on January 19, 1949, and are 6'5'' tall and weigh 190 pounds steal whenever they feel like it? It is much easier for the members of that select group to will the universalization of that maxim. Kantians have tried all sorts of ways of ruling out such gerrymandered maxims from the universalizability test, but there is no generally accepted way of doing so.

Because the main principle evaluates social practices rather than norms or principles or maxims, it avoids the problem entirely. The main principle evaluates social practices by their effects on the equitable promotion of life prospects, not how they are named. A proposal that allows me and a few others to steal at will is not one that would equitably promote life prospects, regardless of how it is named.

It should also be mentioned that, as a practical matter, the expanded original position heuristic eliminates the motivation for gerrymandering maxims. Because those in the expanded original position have no identifying information about themselves, they would have no motivation to agree to any social practice described by one of the gerrymandered maxims. They would be motivated only to agree to practices that would be acceptable to anyone covered by the practice, no matter what their position. This guarantees that no one will even propose a social practice based on a gerrymandered maxim in the expanded original position.

Why Does the Main Principle Endorse the Unconscionability Exception?

In the previous chapter, I explained how the main principle avoided a number of counterexamples that have been thought to be decisive against consequentialist theories of morality or justice. In this section, I explain why the main principle would endorse an unconscionability exception in morality and in law. A similar explanation could be given for the necessity exception.

I begin by evaluating the unconscionability exception in contract law as a substantive practice. The first thing to say is that such an exception would discourage contracts that are exceptionally onerous to one of the parties—for example, contracts of slavery or a long period of indenture servitude. Because discouraging such contracts solves a collective action problem (CAP) for the potential slaves and indentured servants, it would improve their life prospects, while somewhat reducing the life prospects of potential masters. However, because slaves and long-term indentured servants would have very low life prospects, equity requires giving extra weight to improvements in their life prospects. Thus, there is no question that the unconscionability exception would be favored as a substantive legal practice by the main principle, so long as it does not have other undesirable side effects.

It avoids undesirable side effects for two reasons: First, the exception is narrowly drawn, so it would not make contract law unstable; second, it is drawn in such a way as to avoid the potential costs of conflicting judicial interpretations. What is interesting about the unconscionability exception is that it has done a good job of voiding exploitative contracts even though no one has ever successfully explained what it is about a contract that makes it *unconscionable*. Even without a definition, the concept of "shocking the conscience" is able to pass the substantive evaluation test of the main principle, because very few contracts will be affected (narrowness) and the standard is one that does not require fine discriminations on which there would be expected to be lots of reasonable disagreement. By contrast, imagine the variability in judicial decisions if the standard were "troubles the conscience." A "shocks the conscience" standard avoids this problem and so can be endorsed as a substantive practice by the main principle.

What about the evaluation of the exception as an implementation practice? When introduced by an adjudication in a common law system, there would be no coordination problem. In a common law system, everyone knows to look to decisions of the highest level appellate court to determine what the law is. A decision from that court endorsing an unconscionability exception would make it part of the law. This way of implementing the exception easily earns the endorsement of the main principle.

What about the unconscionability exception in morality? The substantive evaluation of the practice is much the same as the substantive evaluation of the corresponding legal practice. But implementation is a different matter. The highest appellate court in a common law legal system solves the coordination

problem because what that court holds is the law. But, as the example of Tom and the music lovers' CAP illustrated, an individual cannot solve the coordination problem for moral norms by acting unilaterally. How then could you be justified under the main principle in unilaterally deciding to escape your servitude, even though you voluntarily agreed to perpetual slavery with Marie in order to obtain her cure?

The answer is that the main principle does not endorse solving this coordination problem. The fact that the practice of slaves escaping from their masters will mean that people will be able to place less reliance on slavery agreements in the future is not a problem, when evaluated by the main principle. Although solving coordination problems is one of the primary ways of equitably promoting well-being, not all successful coordination equitably promotes well-being. There is no danger that runaway slaves will undermine the stability of agreements or promises generally, so no matter how many slaves unilaterally void their slavery agreements and try to escape, their decisions will not generate a coordination problem that the main principle would endorse solving. So the main principle would endorse an unconscionability exception to the bindingness of voluntary agreements.

Is the Main Principle a *Consequentialist* Principle of Human Rights?

The main principle is a principle for explaining when a change in moral or legal practices is an improvement. Suppose I am right that it explains why, once a moral tradition has crossed the consequentialist threshold, any moral or legal system would be improved by guaranteeing the human rights on my list. Does that make my account of human rights consequentialist? Yes and no.

Because the main principle applies only to changes in primary ground-level moral practices, it does not explain the requirements on secondary moral practices, those having to do with enforcement. As I have mentioned, I just assume that there is some kind of proportionality constraint on enforcement of rights. Though I am sympathetic to the idea of providing a consequentialist explanation of the proportionality constraint, I do not try to provide one here. For this reason, I cannot claim to have a consequentialist account of the content of secondary ground-level moral norms or principles.

Also, because the main principle explains only the moral appropriateness of *changing* a system of ground-level moral practices to include guarantees of human rights, it does not provide a complete account of the motives for complying with human rights norms, unless genuine moral motivation comes into play only when ground-level moral norms pass the consequentialist threshold and are endorsed by the main principle. This is an interesting idea, but I have not tried to defend it here.

What I believe is that the main principle provides a consequentialist account of the *content* of human rights norms, because, regardless of what ground-level moral practices one starts with (e.g., libertarian natural rights), the main principle will not endorse those practices unless there is no relevant alternative that does a better job of equitably promoting life prospects. This means that there is nothing sacrosanct about any of the libertarian natural rights (or any other starting point for moral improvement). In coming chapters, I explain how the requirement of equitably promoting life prospects explains the moral appropriateness of changes to recognize the 14 ground-level primary human rights norms on my list. Because the main principle explains the content of those norms, my account of the *content* of human rights norms is consequentialist.

My Own Fudge Factors

I am critical of Mill and Rawls for their use of fudge factors in their theories. What about my own account? It seems full of fudge factors, most prominently, the concepts of well-being and equity, neither of which I define. This is an objection that I return to in chapter 14.

Conclusion

Plato proposed his beehive society in *Republic* as the ideal human society that would equitably promote human well-being. Plato's beehive society cannot be ruled out *a priori*. It requires an empirical inquiry to determine the best kind of society for human beings. Human societies have been engaged in that inquiry for thousands of years. We have discovered that Plato was mistaken. The best way of equitably promoting life prospects for human beings is to establish the human rights that form the constitutional framework of an *experimental society*. In coming chapters, I explain not only why constitutional human rights guarantees, but even more fundamentally, moral practices that take those human rights guarantees as central, are so important to equitably promoting life prospects.

FIVE

The Two Deepest Mysteries in Moral Philosophy

> The object on which we fixed our eyes in the establishment of our state was not the exceptional happiness of any one class but the greatest possible happiness of the city as a whole.
>
> —Plato

> The form of government is best in which every man, however he is, can act best and live happily.
>
> —Aristotle

In this chapter I use the main principle to outline solutions to the two deepest mysteries in moral philosophy. Showing how to solve these two mysteries sets the stage for my consequentialist account of human rights in subsequent chapters.

The two deepest mysteries in moral philosophy are mysteries about the metaphysics and epistemology of objective moral values or objective moral truths. The metaphysical mystery is to understand how moral properties relate to the naturalistic properties of things we can perceive. Western moral philosophy begins with Socrates' discovery in the *Euthyphro* that we have no definition for moral rightness and wrongness. Western political philosophy begins with Socrates' recognition in *Republic* that moral norms, even the norm to give others their due, have exceptions. In the *Nichomachean Ethics*, Aristotle observed that, except for those that were trivially true (e.g., It's wrong to engage in wrongful killing), all moral norms and principles have exceptions (II, 6, 1107a, 9–14) and legal norms do, too (V, 9, 1137b, 11–12).

Hume [1740] turned Aristotle's observation into an indictment of moral judgments. If there are no naturalistic sufficient conditions for moral rightness or wrongness, then it is not possible to deduce an *ought* from an *is*—that is, to deduce a statement about rightness or wrongness from a naturalistic description. Hume thought that this was because there were no objective moral properties. Moral judgments were simply expressions of responses in us.

G. E. Moore (1903) tried to give an objective spin to Hume's conclusion. If there are no naturalistic necessary and sufficient conditions for moral terms, then, he concluded, objective moral properties or values must be nonnatural. Moore thought of himself as a defender of objective moral properties, but

because neither he nor anyone else has ever provided a remotely plausible metaphysics for such nonnatural properties, the most significant effect of his arguments has been to provide support for antirealists (e.g., J. Mackie 1977 and Gibbard 1990).

Notice, however, that to say that there are no naturalistic sufficient conditions or naturalistic necessary and sufficient conditions for moral terms is just a fancy way of saying that substantive ground-level moral principles and norms always or typically have exceptions. To solve the metaphysical mystery, it is necessary to explain why ground-level moral principles and norms typically have exceptions and to do so in a way that dissipates the mystery about their metaphysical status.[1] This I will do shortly.

Like the metaphysical mystery, the epistemological mystery about moral judgments arises within a naturalistic world view. If human beings are the product of a blind evolutionary process, how could human beings ever have knowledge or even reliable beliefs about objective moral truths or objective moral values? There are two aspects to the puzzle: First, it is hard to see how objective moral standards could have played any causal role in a blind evolutionary process. If they played no causal role in the evolution of our moral faculties, it is extremely implausible to think that they are playing a causal role in our exercise of them—that is, when we actually make moral judgments. So if our model of knowledge requires causal interaction with what we have knowledge of, then moral knowledge would be totally mysterious.

Harman (1977) generalizes this argument to conclude that objective moral properties play no role in explaining our moral beliefs. Here is the main idea of his argument: If we assume that we are the products of a blind evolutionary process, then there is a complete *explanation* of the moral beliefs that we have that makes no reference to objective moral values or objective moral properties. If this is true and we know it about ourselves, then we must conclude about ourselves that we fail to satisfy a sensitivity condition for a true belief that p to be knowledge or for it to be reliable: that if p were not true we would not believe it.[2] It is because he can see no coherent way that our moral beliefs could satisfy this tracking condition Williams (1985) concludes that the idea of objective moral knowledge is itself incoherent. This is the second aspect of the epistemological puzzle.

As R. Dworkin (1996) points out, the idea that we causally interact with moral properties when we make moral judgments is crazy. Dworkin ridicules the idea by imagining that there were tiny moral particles (morons) given off by acts that enabled us to perceive their moral properties. But Dworkin doesn't help us to understand how evolution could have given us the capacity to be sensitive to moral properties. He leaves it a mystery.[3]

How could we think that evolution has given us the capacity to be sensitive to objective moral truths, if we agree that objective moral properties play no causal role in the evolutionary explanation of how our capacities developed? Is it possible for objective moral truths to play an explanatory role but

not a causal role in our moral beliefs? Is it possible for our moral beliefs to track objective moral truths or, at least, to be sensitive enough to them to make them somewhat reliable, if objective moral truths play no causal role in our moral beliefs? These are all different ways of stating the epistemological mystery.

Although Aristotle did not solve this mystery, in the *Nichomachean Ethics* he provided a framework for solving it with his account of practical wisdom as an ability to make judgments about individual cases even in the absence of exceptionless ground-level norms or principles. I use his framework to outline a solution to the epistemological mystery shortly.

Clearly, the metaphysical mystery and the epistemological mystery are closely related. Any satisfactory solution to the metaphysical mystery must provide us with a metaphysics of objective moral truth that enables us to understand how the judgments of a person with practical wisdom could be sensitive to it. The two mysteries are like equations in two variables that must be solved simultaneously.

Even though it has been more than two thousand years since Aristotle drew the attention of philosophers to the paucity of exceptionless moral norms and principles and since he provided a framework for understanding moral knowledge in terms of practical wisdom, there has been no satisfactory solution to either mystery—that is, no explanation of why moral norms and principles generally have exceptions that resolves the mystery of the metaphysical status of objective moral truth and no explanation of how the ability of the person of practical wisdom to recognize at least some exceptions to substantive moral norms and principles could involve sensitivity to something objective. Because I think that such mysteries get solved by an historical-social process of discovery, rather than immediately addressing the two mysteries directly, I begin with an historical account of the discoveries that have set the stage for solving them.

The Rocky History of Welfare Consequentialism in Political Philosophy

The idea that there is a welfare consequentialist principle of justice is as at least as old as philosophy. I have included as epigraphs to this chapter Plato's and Aristotle's statements of welfare consequentialism. I could have easily added statements from Hobbes, Spinoza, Rousseau, Hume, and Burke. Although nineteenth-century utilitarians tended to claim many of these early consequentialists as their intellectual precursors, it is a mistake to read them as proto-utilitarians, because most, if not all of them, seem to have assumed that justice required not merely promoting overall happiness, but *everyone's* happiness, which is to say that most, if not all, of these early consequentialists seem to have been taking it for granted that justice includes some sort of equity constraint on the distribution of well-being.

If consequentialism about justice has been a dominant theme in the history of philosophy, why hasn't consequentialism carried the day? Why are there so many philosophers who are nonconsequentialists about justice? Few philosophers today go to the extreme of Kant's *anticonsequentialism*, according to which the goodness or badness of consequences has *nothing* to do with justice. As Rawls says, to most philosophers anticonsequentialism seems "crazy" (1971, 30). But Rawls makes this remark in order to draw attention to the large open space between consequentialism and anticonsequentialism, space for theories that consider consequences (e.g., well-being and its distribution) as significant but not determinative for justice. Most philosophical theorists of justice seem to locate themselves in this space between consequentialism and anticonsequentialism. They are nonconsequentialists but not anticonsequentialists.

One of the main reasons that consequentialism has not carried the day in theories of justice is that, beginning with Plato's *Republic*, consequentialists have typically thought that their consequentialist principle was a ground-level moral principle that was to be applied by individuals—in Plato's case, by the philosopher-autocrat. Plato thought it was obvious that the best way of equitably promoting well-being in the *Republic* was to give absolute power to a philosopher, who by *a priori* insight into the nature of the good life and the good society, could determine the best way to achieve the greatest happiness of the city as a whole.

The indirect consequentialist case for liberal democracies could not be fully appreciated until the direct consequentialist approach to justice had been thoroughly discredited. Even Rousseau [1762] thought it was obvious that everyone would voluntarily surrender *all* of their natural rights in order to enter into civil society aimed at promoting the well-being of all as part of the general will. For Rousseau, no one had any rights against the general will, because genuine freedom could only be achieved in conformity to the general will. The idea behind this claim had a certain plausibility, because, given that the general will equitably promoted everyone's well-being, those who opposed it would either be failing to promote their own well-being or would be seeking to unfairly promote their own well-being at the expense of the well-being of others. When someone out of ignorance acts so as to harm themselves, there is a tendency to think that their action was not fully free. And the idea that unfairly promoting one's own well-being over the well-being of others is a kind of unfreedom is the guiding idea behind Rawls's conception of full *autonomy* as willingness to cooperate with others on fair terms of social cooperation (1993, 77). Thus, the Rousseauian idea of freedom through participation in the general will is a recognizable one. However, Rousseau's use of the idea almost invited totalitarian interpretations—for example, Rousseau's unfortunate rhetorical claim that someone who opposed the general will should be forced to be free ([1762], I:7).

If the French Revolution had ushered in a humane government that equitably promoted life prospects without the need for individual rights, there

would not have been much for the advocates of individual rights to complain about. Perhaps a few philosophers would have insisted that respect for human rights trumped considerations of well-being. But if they had had to argue that everyone should be willing to accept a significant reduction in their life prospects in order to guarantee individual rights, they would have had a hard sell. What made the case for individual rights compelling was that the totalitarian interpretation of Rousseau's general will led to tragedy, the Reign of Terror in France. Somehow, a government that was based on equitably promoting the life prospects of all became the organ of drastically reducing the life prospects of many.

However, the direct consequentialist approach to justice was not thoroughly discredited until the Marxist revolutions of the twentieth century. When Marx and Engels [1848] observed the amount of misery that private property rights had produced under capitalism, they concluded that individual rights were the problem. They thought that all that was necessary to eliminate misery was to establish a government that would equitably promote well-being directly, unencumbered by individual rights. What they failed to anticipate was that a state that guaranteed no rights would be one that exercised absolute power and that such a state would generate a competition to exercise that power, in which victory would go to the most ruthless.

If one were to characterize the rule of Stalin, Mao, Pol Pot, and Kim Il Sung, it would have to be said that, if they equitably promoted anything, it was misery rather than well-being. Some of the greatest famines in history took place under Stalin, Mao, and Kim Il Sung. Pol Pot did not starve the Cambodian population, but he killed a larger percentage of his country's population than any of the other three. In the communist revolutions, the consequentialist exception to human rights was discredited a second time, a second tragedy. So the experiment with applying the main principle as a ground-level principle was an utter failure.[4]

The experiment with Marxist dictatorships produced a greater appreciation for the problems that a government must solve to equitably promote the well-being of its citizens. In Talbott (2005) I identified two problems: the *reliable feedback problem* and the *appropriate responsiveness problem* (36–38). The former is the problem of designing a government that will receive reliable information about how well the government's policies are promoting the well-being of the citizenry. This is no easy problem for an autocrat to solve, because officials will know that their jobs depend on reporting what the autocrat wants to hear.

The problem of appropriate responsiveness is the problem of designing a government that can be relied upon to respond appropriately to reliable feedback, to use the feedback to improve the extent to which its policies equitably promote the well-being of its citizenry. This problem is illustrated by Sen's (1999) research on famines. China had the greatest famine in human history during 1958–1962. Thirty million starved. Mao himself was willing to starve ten times that number (Chang and Halliday 2005, 439), so no one could think

that he was appropriately responsive to feedback about the extent to which his policies were promoting human well-being. In contrast, Sen (1999) reports democracies with an active opposition and freedom of expression and freedom of the press don't have famines.

The Marxist dictatorships greatly discredited direct consequentialist approaches to justice. But Marx was right that there was a problem with laissez-faire capitalism. What was the problem? The problem was dramatized in the United States by Franklin Roosevelt's confrontation with the Supreme Court over his New Deal legislation for regulating the economy. The Supreme Court had enshrined individual liberty as a value that could not be compromised by considerations of well-being, but the liberty on which they placed such a high value was liberty of contract—the freedom to bargain. In *Capital*, Marx described clearly how freedom of contract produced competition among unskilled laborers that made it inevitable that they would agree to work for subsistence pay in occupations that tended to kill or injure them. Although he did not have the terms to describe it, what Marx was clearly describing was the fact unskilled workers faced a collective action problem (CAP) in which they would all be better off if they were legally prevented from offering to work for less than a statutory minimum wage. The same logic applied to maximum limits on hours of work and to the provision of safe working conditions. However, in the United States, almost all economic regulatory legislation was struck down as unconstitutional by the U.S. Supreme Court on the grounds it was an encumbrance on individual liberty.

It was the Great Depression that discredited this theory of the priority of liberty, simply because the extent of the misery generated by the capitalist economy was so great and there was no reasonable prospect of its being alleviated by the workings of that system. The New Deal introduced extensive economic regulation, including minimum wage and maximum hours legislation, banking legislation, and collective bargaining legislation that we take for granted today, but at the time, it generated a constitutional crisis as the Supreme Court was poised to invalidate most of it. In response, Roosevelt threatened to pack the Supreme Court. As is well known, the Supreme Court "blinked" when Justice Roberts switched sides in *West Coast Hotel Co. v. Parish*,[5] upholding a Washington state minimum wage law and overruling decades of precedents.

This decision represented the repudiation of the libertarian idea, articulated by Kant [1793], that governments should provide a framework for liberty or autonomy but not try to promote well-being with legislation.[6] Of course, it was already well-known that liberty rights can generate collective action problems. Although the term did not exist when he wrote, Hobbes was clearly aware of the logic of a CAP and he realized that unrestricted liberty rights would generate CAPs that, if they were not solved, would make life "solitary, poor, nasty, brutish, and short"([1651], 107). What the Marxist experiments showed was that even if coercive governments can be justified as solutions to collective action problems, it is by no means easy to design a

government that will effectively and equitably solve its citizens' collective action problems.

Why would a Marxist dictatorship, understood as the legal establishment of a ruler who directly applies the main principle, unconstrained by individual rights, fail to satisfy that very principle? The most obvious problem is that such a practice would inevitably generate an unbridled competition for the power to rule without constraint, which is often won by the most ruthless competitor, not by the person with the greatest concern for equitably promoting well-being. Although future generations will learn few positive lessons from the history of communism in Russia or China about how to equitably promote well-being, political scientists will continue to study that history for insight into the dynamics of unconstrained power politics.[7] In my earlier volume (Talbott 2005), I referred to this part of the problem of justifying a benevolent autocracy on consequentialist grounds as the *benevolent motivation problem*, itself a special case of the *appropriate responsiveness problem*. This is the problem of designing a government that aims at equitably promoting the well-being of its citizens. No autocracy has ever solved this problem, because autocracies select for leaders who are ruthless power seekers. Ruthless power seekers do not tend to place the interests of their people above their own interests.

But there is a second problem that made it inevitable that Marxist dictatorships would fail the consequentialist test. Even if they are well-intentioned, not even the smartest human beings are very good at predicting what practices will equitably promote well-being, unless the practices have already been tried and already have an established track record. In the first half of the twentieth century, Western intellectuals were easily persuaded that replacing private ownership of the means of production with government ownership would be a good thing, when in fact its effects were disastrous. In chapter 9, I discuss more fully the role of property rights in promoting well-being. Because even well-intentioned human beings are not very good at knowing how to equitably promote well-being, the only way to make progress is by a process of trial and error. Such a process depends on the government's receiving reliable feedback on the extent to which its policies have been successful in promoting well-being (the *reliable feedback problem*). In Talbott (2005), I explained the role of autonomy rights and democratic rights in solving this problem.

The Replacement of Institutions Dominated by One or a Few Individuals with Broader Social Processes

Human beings are not very reliable in their evaluations of untried social practices under the main principle. Indeed, they are so unreliable that, for the reasons discussed above, a change to a government that explicitly applies the main principle to justify its legislation would itself not be endorsed by

the main principle. This is the paradox of direct consequentialism. It is useful to take some time to investigate this paradox more carefully.

If we can't rely on individual autocrats to promote justice, how in the world could it ever be promoted? We now have enough of a history of promoting justice to be able to outline an answer to that question. Justice is best promoted by social processes that reflect group judgments rather than social systems dominated by one or a few individuals. Consider, for example, the following processes: trial by jury, common law adjudication and judicial review, market economies, the process of scientific publication and peer review, the free give-and-take of opinion that is now legally protected by rights to freedom of expression, and democratic political processes.[8] None of these processes has virtues that are evident *a priori*, as illustrated by the fact that Plato would have opposed every one of them. Even those, like trial by jury and democratic political processes, that have come to be formally enshrined in constitutions, had precursors that were not the product of intentional design.

Also, the nature of the processes themselves has evolved—for example, unanimity requirements on jury verdicts have been weakened, standards of judicial review have evolved as a result of the process of judicial review itself, market economies have been modified with various kinds of government regulation, the process of scientific publication and peer review is in the process of changing to require disclosure of funding sources for research and of the financial interests of researchers in their research, the norms for respectful discourse evolve over time as a result of the free give-and-take of opinion about them, and democracies often revise the requirements for voting. Of course, these processes are made up of individuals. But none of these processes came into existence by the intentional design of a single individual, and in none of them does a single individual have decisive authority to determine their output.

If autocrats can't be relied upon to produce changes in social practices endorsed by the main principle, then the only hope for us is that there are some social processes that have the power to do so. The idea that social processes could be the source of improvements in knowledge or in morality or justice was startlingly new when first introduced by Hegel [1821]. Hegel could see that there was progress in history, but not even Hegel realized what was driving the progress. Mill was one of the first to realize that in every area of inquiry and every human activity, progress was driven by the free give-and-take of opinion and by freedom to try new things. And so when Mill sat down to write the first indirect consequentialist defense of robust, inalienable individual rights, the first one on the list was the right to freedom of thought and discussion ([1859], chap. 2). I return to the importance of the Millian epistemology for human rights in chapter 7.

Science remains the most powerful example of the epistemological potential of social processes, but it is now recognized that progress in every area of knowledge depends on social processes. Even in mathematics, it is the

community of mathematicians that determines whether a proffered proof is fallacious or not.

In the political realm, it is democratic processes that have played the biggest role in accelerating the process of equitably promoting life prospects. And thus it is that the main principle endorses democratic forms of government, as I discuss more fully in chapter 10.

Why Doesn't the Main Principle Endorse Judicial Application of the Main Principle?

Consider the unconscionability exception again. In chapter 4 I discussed why it would be endorsed by the main principle as an exception in both law and morality. But why should the main principle endorse exceptions in this piecemeal fashion? Why doesn't it just endorse itself as a ground-level legal and moral principle? I focus on its role in the law, because parallel considerations apply to its role in morality. To answer my question, it is useful to consider two kinds of ground-level legal principles and explaining why the main principle does not endorse either of them.

First, let's consider why the main principle doesn't endorse the following rule of jurisprudence: Enforce established law except when an exception would equitably promote life prospects. This principle is not even an application of the main principle, because it would have judges make exceptions on a case by case basis. If judges applied this principle, in an action for petty theft involving an impoverished thief and a wealthy victim, judges would typically be required to rule for the thief, because the relatively small loss would have a negligible effect on the wealthy person's life prospects, but would significantly enhance the life prospects of the poor thief and would also promote equity. Of course, such an exception would have even more disastrous results than the Marxist revolutions, because it would eliminate property rights even more thoroughly than they did. Thus, it would not be endorsed by the main principle.

This suggests that the problem with the first statement of the rule was that it applied to acts rather than to practices. So let's consider a second principle: Enforce established law except when an exception, considered as a practice, would equitably promote life prospects. This principle would never endorse an exception to petty theft laws to favor impoverished thieves over wealthy victims. But there would still be a serious problem that would prevent this practice itself from being endorsed by the main principle.

Recall that the main principle evaluates a proposed exception as a substantive practice, on the assumption that it is successfully implemented, and as an implementation practice. One of the advantages of discussing legal practices in a common law tradition is that when judges follow the norms for announcing their decisions, there is usually no coordination problem generated by implementation. However, this would not be true if judges applied

the main principle in their decisions. That practice would generate multiple coordination problems.

Imagine that each judge applies the main principle to each case to decide when to make an exception to the existing law. We would expect lots of reasonable disagreement among different judges about the extent to which different practices would equitably promote well-being. This would be expected to lead to lots of conflicting judgments and, thus, to a kind of legal parochialism, in which each judge's law would be whatever he said it was. If different judges had different laws, this would be a new kind of coordination problem.

Even if different judges' opinions did not conflict with one another, so that the law everywhere was the same and there would be no difficulty in determining what the law *is*, there would still be a serious problem of determining what the law *will be*. If judges made exceptions to the law whenever they thought that the exception, as a practice, would equitably promote life prospects, this would make the law too unstable to be a basis for reliable predictions about what the law would be in the future. Notice that this is true even if every individual change alone would equitably promote life prospects. So the main principle would not endorse itself as a principle of adjudication.

The most important reason that the main principle would not endorse itself as a rule of jurisprudence is that it would make each individual judge a legislator. But on most matters, a democratic process will do a much better job of enacting laws that equitably promote life prospects than any individual would do. So the main principle will endorse distinguishing the role of judge from the role of legislator. Even in a common law system in which there is no legislature, the system as a whole will better satisfy the main principle if judges base their decisions on considerations of rights and fairness than if they explicitly apply the main principle to each case.

A Solution to the Metaphysical Mystery: Why There Are Exceptions to All (or Almost All) Ground-Level Moral and Legal Norms and Principles (and Exceptions to the Exceptions, Etc.)

As I have mentioned, it was Aristotle who first drew our attention to the difficulty of formulating exceptionless moral principles and exceptionless laws. In Western philosophy, because reasoning was long thought to be top-down, from principles to particular judgments, the lack of exceptionless principles has been a great scandal—at least to moral realists. Kant [1799] was so sure that moral reasoning required categorical principles and norms that, when presented with exceptions to his principles and norms he responded with denial—for example, by denying that it could ever be right to lie, even if it were necessary to save a life. It was not until the twentieth century that moral realists began to study carefully the logic of moral norms that are defeasible,

not categorical. Ross (1930) was perhaps the first to study the logic of defeasibility in ethics. R. Dworkin (1977) was one of the first to study the defeasibility structure of legal principles.

Among contemporary philosophers, Dancy (2004) has insisted on the importance of the defeasibility structure of moral norms and principles. He has also emphasized the importance of the ability to make judgments about particular cases that cannot be subsumed under any accepted principle or rule. This makes him a *moral particularist*.

If it is true that all interesting moral and legal principles have exceptions—which, of course, implies that there are exceptions to the exceptions, and so forth—this is a surprising fact that leads us to wonder, how is it possible? How is it possible that finite creatures like us could be sensitive to a potential infinity of moral distinctions? Of course, it is true that, as finite creatures, we can be sensitive to only a finite number at any one time. But if it is always possible to add one more, then there is no finite limit. How could this be?

A Model for Defeasible Reasoning: Reference Class Logic

A moral particularist like Dancy can't really explain it. When Dancy himself addresses the puzzle, the best he can do is to point out that defeasibility is not limited to moral and legal reasoning. All reasoning seems to have the same structure (2004, 73–85). So, for example, my seeming to see a red desk can justify me in believing that I am looking at a red desk, though it would not if I realized that the desk was being illuminated by red light.

If we are puzzled about the defeasibility of moral and legal reasoning, it is, in a way, reassuring to notice that all reasoning is defeasible. It takes away some of the worry that moral and legal reasoning are a defective kind of reasoning. However, in another way, the puzzlement is increased, because now we have an even bigger puzzle than the one we started with. Why is all reasoning defeasible? The key to what seems to me to be the most promising explanation is to see that the defeasibility structure of reasoning can be modeled by the logic of statistical reasoning with conditional probabilities.

Consider a nonmoral example. In order to plan for my retirement, I need information on how long I will live. If I try to determine my life expectancy, I find that it varies with the information that I have. Given only that I am a male U.S. resident, it is 74.8 years. Given only the additional information that I am 60 years old, it is 20.36 (i.e., a total of 80.36) years. Given only the additional information that I have never smoked tobacco, it is 24 (i.e., a total of 84) years. Notice that my life expectancy changes whenever I add further relevant information to what is called the *reference class*—that is, the class of people who are like me in the relevant respects for which the statistics are known. Because the probabilities that determine life expectancy are based on a reference class, they are *conditional* probabilities. Conditional probabilities need not approach a limit as more

information is added to the reference class. No matter how much information we have added, it is always possible that there is some further information, which, if added to the reference class, would make a big difference to my life expectancy. For example, if it were true that I had just been diagnosed with colon cancer, adding that information would greatly reduce my life expectancy.

Because the results of reasoning with conditional probabilities are relative to the reference class, I refer to the logic of this sort of reasoning as a *reference class logic*. It is easy to understand why a reference class logic applies to statistical reasoning that explicitly employs conditional probabilities. Why would it apply to other kinds of reasoning that are not explicitly probabilistic? The answer is that most of our reasoning can be *modeled* at the meta-level by statistical reasoning, even when we are not aware of doing any reasoning. For example, you are probably very reliable at distinguishing red things from white things in normal lighting. But if you find out that red light is being used, you would probably adjust your degrees of confidence in your judgments about things that appeared to be light red. This change in your degrees of confidence would not be due to any consciously probabilistic reasoning, but it can be modeled by reasoning with conditional probabilities: The conditional probability of your correctly identifying a light red object, given that the light is red, is less than the conditional probability of your correctly identifying a light red object in normal lighting, because in red light white objects look light red.

The general conclusion to draw is that any reasoning that is *modeled by* conditional probabilities will have a reference class logic and be defeasible (if the conditional probabilities involved are neither zero nor one). What about moral and legal reasoning? Both kinds are defeasible. Moral and legal reasoning are not explicitly probabilistic. Could it be that they are *modeled by* reasoning with conditional probabilities? I think the answer is yes.

A Probabilistic Meta-Level Model for Moral and Legal Reasoning

To understand the reference class logic of moral and legal reasoning, recall that the main principle evaluates practices on the basis of their effect on life prospects. At every stage of life, life prospects are defined by a probability distribution. Life expectancy is also defined by a probability distribution. So reasoning about life prospects will exhibit the same reference class structure as reasoning about life expectancy. However, reasoning about life prospects is explicitly statistical, just like reasoning about life expectancy. To understand moral and legal reasoning, we have to understand why it has the reference class logic of reasoning about life prospects, even though it is not explicitly about life prospects and is not explicitly statistical. In other words, we have to understand why, though it is not

explicitly *about* life prospects, it can be *modeled by* meta-theoretical reasoning about life prospects.

Consider the example of life expectancy again. On any plausible consequentialist principle, increases or decreases in life expectancy will be morally relevant to the evaluation of alternative acts. Thus, any plausible consequentialist principle will permit exceptions to the norm of truth telling when telling the truth would have a drastic effect on people's life expectancy. That is why Kant was wrong to think that it could never be wrong to tell the truth.

Because the concept of life expectancy itself has a reference class logic, this fact alone provides the consequentialist with some reason for expecting moral reasoning to have a reference class logic. Life prospects include other probabilistic elements in addition to life expectancy. Therefore, any moral reasoning modeled by reasoning about life prospects would be expected to have a reference class logic.

It is important to emphasize that moral reasoning might be *modeled by* reasoning about life prospects even if the reasoning itself is reasoning about what is right or fair, not about which practices would best promote life prospects. Moral heuristics such as the Golden Rule or the expanded original position provide a framework for moral reasoning that does not require explicit reasoning about life prospects, but which might well be *modeled by* reasoning about life prospects.

Consider an example. When we consider a norm against lying, in our ground-level moral reasoning, we see both positive and negative reasons to support it. On the positive side, when we think of compliance with the norm, we think of the possibility of mutual trust and cooperative activity, which have a powerful moral allure. On the negative side, when we think of deception, we are struck by how the deceiver "uses" the other person for his own ends. We see the disappointment that the "used" person feels when she discovers that she has been duped into acting to promote the other person's interests at the expense of her own.

I hope it is clear how to translate this moral reasoning into a model in terms of promoting life prospects. Practices that promote cooperation and trust greatly promote life prospects. Being "used" for another person's ends greatly generally diminishes one's life prospects. Indeed, "using" another in this way is easily understood as a way of free riding in a collective action problem. Of course, the very definition of a collective action problem requires reference to expected utility or some probabilistic measure of life prospects.

Or consider a norm of keeping one's promises and performing on one's voluntary agreements. At the ground level, breaking one's promise or failing to perform on an agreement strikes us as a breach of trust—as a failure of respect for the other. At the meta-theoretic level, the practice of promising and performing on one's agreements provides a framework for *productive exchanges*—that is, exchanges that promote the life prospects of all parties.

Again, those who fail to keep a promise or fail to perform on an agreement typically can be seen to be free riding on the cooperation of others.

If we think of moral reasoning as having a reference class logic, then it is easy to understand how any norm can have exceptions. However, it must be acknowledged that the main principle supports a fairly strong presumption against exceptions to moral norms. There are two reasons for this. First, most exceptions to moral norms have the potential to disrupt coordination. Because one of the main generic benefits of moral codes is their coordinating function, there will be an initial presumption that exceptions will reduce life prospects. Second, for a moral code to perform its coordinating function, it must be fairly easy for people to learn it. A proliferation of exceptions, each individually of marginal benefit, might jointly make the code too unwieldy to be an effective coordinator of expectations and behavior.

Nonetheless, at the ground level, the intuitive appeal of some exceptions is almost irresistible. Consider again the norm of promise keeping. It would be crazy never to allow exceptions to this norm. The explanation for this at the meta-level is that there are simply too many kinds of cases in which the practice of allowing the relevant kind of exception to the promise-keeping norm will promote life prospects or, to translate this into ground-level moral reasoning, will seem unfair. We have already discussed exceptions for actual consent, necessity, and unconscionability. It may seem that it would not be necessary to include an exception for coercion or duress, because it would seem that only voluntary promises made without coercion or duress would be binding. However, I would not be surprised if there have been traditions that at one time held all promises to be binding. Hobbes, for one, argued that promises based on fear were voluntary and binding (1651, chap. 14). My impression is that at least some moral traditions begin with a fairly strong standard of strict liability and that the idea of fault is a later development. I discuss this development in the next chapter.

There are many other bases for exceptions to promises and voluntary agreements. Consider, for example, mistakes. Suppose you pay me for 100 widgets and I promise to deliver them on June 31. Neither of us remembers that June has only 30 days. On July 1 you call me to ask when I am going to deliver the widgets. I tell you that I am not required to deliver them until June 31, so you'll never get them. But I'm keeping your money.

On a strict construction of the agreement, I have not failed to perform. But no one would think it was fair that I keep your money and never deliver the widgets. In the law this would be considered unjust enrichment. What is the corresponding model-theoretic reasoning? At the model-theoretic level, we should expect there to be some impetus for exceptions to rule out unproductive exchanges, because of their adverse effects on the life prospects of one party to the agreement. In addition, there is an extra reason for making exceptions to rule out unproductive exchanges: incentive effects. The feeling that it would not be fair to allow me to keep your payment for the widgets without delivering you any widgets tracks the

model-theoretic fact that rewarding me for the mistake on the date will provide an incentive for promisors and promisees to try to induce the other party to make a mistake. This threatens to significantly reduce the benefits of promising. I return to this topic when I discuss the doctrine of *caveat emptor* in chapter 9.

Consider one more kind of exception: impossibility. Suppose I grow a unique species of tomato, not available for sale in stores. You trade me some of your special beans today for my promise to give you some of my special tomatoes next month. Then my tomato crop is wiped out. I can't give you any of my unique species of tomatoes. Clearly, an exception to keeping my promise is allowed here. However, in this case and in almost all cases in which there is an exception to keeping one's promise or performing on an agreement, fairness requires some compensation to you. I am not permitted to just say "Tough luck" and walk away. Why is compensation required by fairness?

At the meta-level, model theoretic reasoning favors the compensation requirement because it makes it probable that the life prospects of both the promisor and the promisee are increased by the practice. Also, it maintains the structure of a productive exchange, and thereby avoids creating incentives for the promisor to act so as to make performance impossible.

In this section I have focused on moral reasoning. The case for a meta-level model-theoretic analysis of legal reasoning is even stronger. In the law of contracts alone, the legal exceptions to the general rule that a contract is enforced as written include not only what I have referred to as necessity, unconscionability, mistake, and impossibility exceptions, but also exceptions for impracticability, frustration, illegality, misrepresentation, and, for commercial contracts, implied warranties of fitness and merchantability (implied because they are enforced even if not explicitly stated in the contract). And then there is bankruptcy, which itself has a whole variety exceptions to contractual claims of creditors. I discuss these exceptions more fully in chapter 9. Here it suffices to say that almost all of them are examples of exceptions to avoid unproductive exchanges and to avoid incentives for unproductive exchanges. Of the items on the list, the two exceptions to this rule are the illegality exception and the bankruptcy exceptions. Enforcing illegal agreements would obviously reduce life prospects, though in a different way than unproductive exchanges do. Bankruptcy is so clearly a consequentialist exception to contracts that it hardly needs to be argued. I say something more about bankruptcy in chapter 9.

The Final Element of the Solution to the Metaphysical Mystery

We are close to being able to explain the defeasibility of ground-level moral reasoning. To finish the explanation, we must consider only one final question. Why not replace our ground-level moral reasoning with the explicitly probabilistic meta-level reasoning based on the main principle? That is, why

can't the main principle itself serve as an exceptionless principle of ground-level moral reasoning?

The answer has already been given. Because of the paradox of direct consequentialism, the main principle ranks our actual practice of moral reasoning based on defeasible principles of right and wrong, of fairness, and of rights as superior to a practice of moral reasoning based explicitly on the main principle.

If I am right that moral and legal reasoning really have a reference class logic, then it should be possible to proliferate exceptions and exceptions to exceptions without limit. Is this true? Can we really do it? Consider the moral practice of promise keeping. If I promise to meet you at noon at my office, I should be there. But if on the way to my office I have an opportunity to save a life, I am excused from my promise, even if I have no way of contacting you to obtain your consent. But what if something happens to make me realize that if I keep my promise, I will save three lives? Then it would seem that I should keep the appointment. It might seem as though there is a simple principle to cover these exceptions: to do whatever will save the most lives. But this is not correct. What if the one person whose life would be saved if I break my promise is my child and the three people whose lives would be saved if I keep my promise are murderers? This simple example shows how easy it is to proliferate exceptions and exceptions to the exceptions.

Now consider the law of contracts. I have listed 11 exceptions to the general rule of enforcing contracts as written, but if we were to look at them closely, we would find that almost all of them are labels for a more complex logical structure that includes multiple exceptions and exceptions to the exceptions. This is exactly the structure that we would expect if the logic is a reference class logic. Of course, by attributing a reference class logic to the law of contracts, I am implicitly committed to thinking that there is a potential for further improvements in the law of contracts and that there will always be a potential for further improvements. Isn't this almost certainly true?

So the solution to the mystery of the metaphysics of moral judgment is not that moral concepts pick out some kind of weird nonnatural property. It is that they have an implicitly *probabilistic* metaphysics. Consider the probabilistic concept of life expectancy again. No matter how much information about me that I use to determine my life expectancy, it is always possible that there is further information that, if I had it, would lead me to change the determination of my life expectancy. Moral judgments have the same kind of logical relation to naturalistic descriptions. No matter how much information I use to determine what I ought to do in a given situation, it is always possible that there is additional information that, if I had it, would change my determination of what I ought to do. The metaphysics of moral judgment, whether ground level or meta-theoretical, is no weirder than the metaphysics of life expectancy.

A Solution to the Epistemological Mystery

What about the epistemological mystery? Even if the main principle is an exceptionless moral meta-principle that explains the reference class logic of ground-level moral reasoning, how could a blind evolutionary process have developed beings who were sensitive to such truths? If the probabilistic structure of the main principle solves the metaphysical mystery, then the remaining question is how we human beings could have developed a sensitivity and responsiveness to that principle.

One way that we might have become sensitive to moral truth would have been for us to develop *a priori* insight into moral truths. Many philosophers have claimed to have such insight into moral truths, but the mere fact that they all disagree so much about the content of the insight shows us that it could not be very reliable. So if we are sensitive to moral truth, the sensitivity must have some other source.

One of the most popular arguments for the existence of God is that God could be part of an explanation of how human beings became sensitive to objective moral truths. God simply implanted the sensitivity in us by giving us a conscience. The God hypothesis solves the epistemological mystery by denying the assumption that we are the products of a blind evolutionary process.

How can we resolve the mystery if we don't deny that assumption? Rather than answer that question directly, let me begin by noting that, just as the metaphysical mystery was a mystery about all normative judgment and all reasoning, the epistemological mystery is a mystery about all normative judgment. We must wonder how a blind evolutionary process could have made us sensitive to truths about what it is rational to believe or to truths about what it is rational to do in nonmoral contexts, as well as to truths about what we morally ought to do.

My suggestion is that normative truths are solutions to normative problems and that the ability to solve normative problems conferred an evolutionary advantage on our ancestors. So, for example, when we discover that we have inconsistent beliefs, the discovery typically makes us uneasy and motivates us to look for a way to eliminate the inconsistency. Why would we care? Why not simply acknowledge the inconsistency and get on with life? The answer, I believe, is that inconsistent beliefs have the potential to make our actions self-defeating, as my actions would be, for example, if I offered to help you based on the belief that you had helped me in the past and then quickly withdrew my offer based on the belief that you had not.[9] Truths about what it is rational to believe are solutions to this sort of problem. Evolution favored those who were inclined to recognize at least salient problems of this kind (sensitivity) and motivated to solve them (responsiveness).

Similarly, if we discover that we have intransitive preferences, this discovery tends to make us uneasy and motivates us to look for a way to eliminate

the intransitivity. Intransitive preferences often directly lead to self-defeat. For example, if I prefer beans to peas and peas to corn and corn to beans, given a choice of the three of them, I might never be able to choose, or, if I did choose, I would immediately second-guess my choice and make another, with the result that I might never get to eat any of them, even if I would much prefer eating any one of them to not eating any of them. Truths about what it is rational to do or prefer in nonmoral contexts are solutions to this sort of problem. In this context also, evolution favored those who were inclined to recognize at least salient problems of this kind (sensitivity) and motivated to solve them (responsiveness).

What sort of problem are moral practices solutions to? The most important category is collective action problems. In collective action problems, individual rationality leads to collective self-defeat. Moral practices such as practices of promising and truth telling solve collective action problems. The main principle is the principle that ranks moral practices, at least in part, on how well, when evaluated as substantive practices and practices of implementation, they solve such problems.

There can be little doubt that evolution has made us sensitive to at least salient collective action problems and has motivated us to solve them. Evolutionary psychologists have offered striking evidence that human cognition includes a module for cheater detection—or, as I would describe it, a module for detecting free riders in collective action problems (Cosmides and Tooby 1992). This is sensitivity. Other experimenters have shown that human beings are motivated to sanction free riders, even at some cost to themselves (Dawes, Orbell, and Van de Kragt, 1986; Fehr and Gächter, 2000). This is responsiveness. Perhaps surprisingly, primatologists have found evidence of this sort of sensitivity and responsiveness (i.e., some sanctioning of free riders) in other primates (de Waal 2006).

It may seem that if our normative judgments are evolutionary strategies for avoiding self-defeat, then there would be no reason to think that normative truths were universal—that is, that they applied to all rational beings. Normative truths for human beings would be parochial, applying only to beings with our evolutionary history.

Although this is one possibility, it is important to see that there is a coherent alternative. On this alternative, there are objective normative truths that apply to all rational beings, because the problems they are solutions to are potential problems for any rational being. In giving us the resources to solve these problems, evolution would have, inadvertently, made us sensitive to objective normative truths.

How sensitive? Surely our ground-level moral judgments about particular cases, whether individual or collective, are not infallible. In many noncontroversial cases, it is quite plausible that most peoples' ground-level moral judgments track the results of the main principle. But it is also clear that in many cases they do not. This shows that tracking is too demanding a conception of sensitivity for all our moral beliefs.

How sensitive would our ground-level moral judgments have to be to the results of the main principle for us to have any reason to trust them? It seems to me that, at a minimum, collectively, we would have to be sensitive enough that the changes in our moral practices over time would generally be endorsed by the main principle as improvements. If our sensitivity were so low that most of our changes in our moral practices made us more likely to worsen them than to improve them, it would be hard to think that we should place much confidence in them.

Even if it is possible that evolution has made our normative judgments sensitive to normative truths that apply universally to all rational beings, how could we ever have any reason to believe it? My suggestion is that we begin by trying to figure out what the principles are that explain our normative judgments. Then we can consider whether those principles have features that make them contingent on our evolutionary history or whether they seem to be principles that would apply to us regardless of our evolutionary history.[10] In chapter 1, I explained why the utilitarian principle of maximizing overall (expected) utility has the form of a universal rather than a parochial principle of morality.

What about the main principle? The promotion of one's own life prospects would seem to be a goal for any rational being. This would make solutions to collective action problems important for any rational being who might be in collective action problems with other rational beings. Is there any reason to expect a blind evolutionary process to favor a universal standard of equitable division of the benefits and burdens of cooperation in a CAP? I think the answer is clearly no. Evolution has produced human societies that have no problem justifying very inequitable division of the benefits and burdens of cooperation. For example, almost all human societies are patriarchal societies in which males receive a disproportionate share of the benefits of cooperation without shouldering a correspondingly disproportionate share of the burdens. However, until relatively recently, these inequities have not triggered pangs of conscience in men or much resistance in women, so it is hard to think that evolution made either men or women sensitive to this sort of inequity.

But even if evolution did not make human beings sensitive to this sort of inequity, evolution has given us the equipment to become sensitive to it. Because we have the ability to think about things from an impartial point of view—for example, to think of things from the point of view of the expanded original position—we have the ability to critique whatever standards of equity are current in our society. In the expanded original position, the unequal division of the benefits and burdens of cooperation in patriarchal societies would be rejected as unacceptable. I see no reason to think that the reasoning that leads us to this conclusion is parochial in that it applies only to beings with our evolutionary history. Behind the veil of ignorance in the expanded original position, any rational being should be able to recognize the inequity of the division of the benefits and burdens of cooperation in

patriarchal societies and the potential for moral improvement by reducing or eliminating the inequity. Considerations of this kind at least make it plausible to think that the main principle may be a universal principle of moral improvement.

What Is Practical Wisdom?

If moral and legal reasoning have a reference class logic, then moral and legal codes may be infinitely improvable. But how are we actually able to improve them? The simple policy of designating a single authority to explicitly apply the main principle to decide how to improve them has been tried, with disastrous consequences. This is the paradox of direct consequentialism.

At the other extreme would be a blind invisible hand process that produced improvements even though no one involved in the process had any idea how to improve the moral or legal code or any intention of doing so. Some improvements in the legal code seem to happen this way. If members of a democratic legislature simply aim at improving the life prospects of a majority of their constituents and do a good job of it, changes in the legal code would at least benefit a large portion of the population. If different laws benefited different majorities, it is possible that the aggregate of changes would benefit everyone.

However, there is one kind of improvement that would rarely, if ever, be made by such a blind process—changes that promote equity for an oppressed or stigmatized minority. If all improvements in legal codes were the result of such a blind process, legal segregation would still exist in the southern United States, apartheid would still exist in South Africa, and there would be no gay and lesbian rights anywhere in the world. Many of the most important improvements in moral and legal codes depend on there being enough people who can recognize potential moral improvements and who care enough about correcting them to make at least small sacrifices to do so. This kind of judgment requires the ability to adopt an impartial viewpoint and a capacity for empathic identification. If people did not have these abilities, the process of moral and legal improvement would not only be retarded, some kinds of improvement would be impossible. Many potential improvements would never be made.

Even though there are social processes that are more reliable than any single individual in improving moral and legal codes, those social processes would be not be effective if they did not contain individuals who had the ability to recognize improvements. This is surprising, given that most cultures regard their moral norms as exceptionless rules obtained from an infallible authority. So, for example, even though almost all of the world's major religions have historically endorsed some kind of slavery or caste system, enough people in each of these traditions have come to recognize

that eliminating slavery and caste systems would be a moral improvement that now almost all of the world's religions reject them.

How is this possible? How can training in moral norms, typically moral norms regarded as exceptionless, produce people with the ability to recognize exceptions to those very norms? We now have the elements to put together an answer to that question. Once a society's norms closely enough approximate those that would be endorsed by the main principle, training in the norms produces people who are not just blind rule followers, but who can respond to the *spirit* as well as the letter of the rules. When a society has reached this stage, I say that it has passed the *consequentialist threshold*. Once a society has passed this threshold, the society's moral training produces people who are responsive to the spirit of the rules, who have the ability to recognize the potential for at least some improvements to them. I think this is what Aristotle was referring to by *practical wisdom*.

Although Aristotle insisted that morality could not be codified in exceptionless rules, he also insisted that practical wisdom involved rules in some way—in particular, that it involved "the *presence* of the the right rule" (NE VI 13, 1144b 23–29), but not as Socrates thought or, I would add, as Kant thought. That is because Socrates and Kant thought that morality involved explicitly applying a rule or principle given by reason. Aristotle was pointing to an alternative role for principles—as implicitly guiding the judgment of the person of practical wisdom. I believe that Aristotle was right about the role of principles or rules in practical wisdom. Aristotle never did try to identify the rule or rules involved in practical wisdom. I think we are now in a position to do so. The person of practical wisdom has an implicit sensitivity and responsiveness to the main principle: a sensitivity that enables the person of practical wisdom to recognize exceptions to whatever the existing moral and legal norms may be and a responsiveness that motivates her to act on the judgments produced by her sensitivity.[11]

Although the moral practices of a society typically include a moral code, no society's moral practices could be adequate if they produced only rigid rule followers. As feminist advocates of the care perspective in ethics (e.g., Gilligan 1982; Noddings 1984) have emphasized, ethics crucially involves a kind of emotional responsiveness. Even to recognize the morally relevant features of a situation often requires empathic understanding. This kind of responsiveness precedes the application of rules, so it cannot be guided by rules.

In addition, good moral training develops not only a sensitivity to morally significant factors in a situation, but, paradoxically, it also develops an insensitivity to morally discreditable ones. This is the phenomenon of "silencing" (e.g., McDowell 1978). Consider, for example, Scanlon's example of someone you think of as your friend who knows that you need a kidney transplant, so he decides to find someone whose organs are compatible with yours and then kill that person and harvest a kidney for you (1998, 164). Scanlon uses the example to show how even the seemingly nonmoral concept of friend has

moral content. I use the example to illustrate how the moral content typically cannot be captured by rules. We would expect a good friend to be eager to help you obtain an organ donor. Your friend would probably want to be tested for compatibility to find out if she could donate herself. But it would be disturbing if she were seriously considering trying to find "involuntary" donors. Indeed, it is a little disturbing if the thought even occurs to her. We expect such alternatives not even to suggest themselves.

Obviously, this sort of silencing of morally discreditable reasons cannot be achieved by following a ground-level rule. A rule for silencing considerations of a certain kind would be as self-defeating as trying to prevent yourself from thinking of elephants by reminding yourself not to think of elephants. And yet silencing is an important part of the moral sensitivity that is produced by moral practices endorsed by the main principle. I discuss an example of silencing in the next chapter.

In sum, although the main principle would not endorse adopting a system in which a single individual had the power to change the moral norms or the laws of a society whenever that one individual thought the change would be an improvement, improvements in the moral practices and laws of a society would be haphazard at best, were it not for the fact that a society's moral practices produce individuals whose implicit sensitivity to the main principle enables them to recognize at least some ways of improving those very practices—that is, produce individuals with practical wisdom. In a society that guarantees human rights, everyone has the potential to acquire this kind of sensitivity.

Leveling the Playing Field between the Consequentialist and the Nonconsequentialist

Does my solution to the two mysteries rule out ethical nonconsequentialism? Surely not. But it does remove the major advantage that nonconsequentialism seemed to have over consequentialism in moral theory. Nonconsequentialism's main apparent advantage was that ground-level moral norms and principles are nonconsequentialist. If those ground-level moral principles and norms all have exceptions, then the focus shifts from the ground level to the meta-level. At the meta-level, there is no reason to presume that the relevant principles are nonconsequentialist.

The reference class logic of moral reasoning gives us good reason to believe that the relevant meta-principles are probabilistic. Almost surely, they involve some probabilistic measure of well-being, such as life prospects. There is also good reason to think that they involve some concept of equity.

This is not enough to rule out most forms of nonconsequentialism, because, with few exceptions (e.g., Kant), almost all forms of nonconsequentialism allow for exceptions to ground-level principles and almost all of them hold that effects on the equitable distribution of life prospects are morally relevant

in evaluating a ground-level moral norm or principle. The nonconsequentialist need only insist that the true meta-level principles rank alternative moral practices on other grounds *in addition* to their contribution to equitably promoting life prospects.

It would be premature to take up this issue now, because we have yet to consider how well the main principle is able to explain the moral appropriateness of our ground-level moral judgments and reasoning, when they are appropriate. That is the burden that I take up in the following chapters.

Comparison to Dworkin's Conception of Law

It is useful to compare my consequentialist meta-theoretical explanation of the moral appropriateness of changes in the law with an influential nonconsequentialist account, R. Dworkin's account in *Law's Empire*. In *Law's Empire*, Dworkin was most interested in cases in which judges establish new legal precedent in civil cases, often by overruling prior precedents.[12] Here I focus my discussion on such cases. Dworkin believes that, in such cases, judges should be understood not as making new law, but as interpreting existing law. The "correct" decision is the one that best interprets existing law. This has the paradoxical consequence that, on Dworkin's account, in making new precedents or overruling prior precedents, judges are simply trying to correctly discern what the law *is*, rather than what it *should be*.

The air of paradox is somewhat attenuated when Dworkin explains why he thinks that most philosophers have misunderstood what is involved in determining what the law *is*. For Dworkin, determining what the law is is *constructive*—that is, "a matter of imposing purpose on an object or practice in order to make of it the best possible example of the form or genre to which it is taken to belong" (1986, 52). Making the law the best possible example of the form to which it belongs can involve what would ordinarily be understood as allowing for making changes in the law. Why didn't Dworkin just say that judges have a duty to improve the law when they can do so in ways that don't generate coordination problems?

The answer is that Dworkin thinks that judges are constrained to make their decisions fit with past decisions. This is law as integrity. Dworkin contrasts his theory of law as integrity with a pragmatic theory that holds that judges should make the law the best that it can be without regard for past decisions.

This is not the place for an extended critique of Dworkin's theory of law as integrity. However, there are two points of contact with my account that provide a useful test of my consequentialist account against his nonconsequentialist one. First, even if the goal is to have a legal system that is as good as it can be, a legal system that achieves that goal might not be (and I believe almost surely is not) one in which judges interpret laws in a way that aims at that goal. It may well be that the best way for judges to make the legal system

the best it can be is by aiming at something else. Call this an example of the *paradox of direct bestness*. This objection cuts two ways. It implies that if the goal of the legal system is integrity, then it might be that individual judges should aim at something else. More significantly, it implies that even if individual judges aim at integrity, the goal might be something else—for example, a legal practice that equitably promotes life prospects.

The second issue raised by Dworkin's account is an important explanatory issue that can help to adjudicate between my consequentialist account of improvements in the law and the various nonconsequentialist accounts, including Dworkin's. The issue is this: Which kind of account can best explain the moral appropriateness of an apparent retroactivity in the application of appellate decisions overruling prior precedents in civil cases to the parties to the appeal? Consider *Henningsen v. Bloomfield Motors, Inc.*[13] In that case, Mr. Henningsen bought a new Plymouth as a gift for his wife. Shortly thereafter, Mrs. Henningsen totaled the new Plymouth. While she was driving, the steering mechanism failed, she lost control of the vehicle, and it crashed into a wall. In order to hold the manufacturer liable for damages, the New Jersey Supreme Court had to void a contract provision that limited the manufacturer's liability to replacement of the defective part, impose liability without fault, and extend that liability to Mrs. Henningsen, even though it was her husband who had purchased the car. The *Henningsen* court overruled prior precedents to reach this result.

The most natural way of describing the problem is to say that in the *Henningsen* case, the appellate court was making new law and applying it retroactively to an incident that occurred 5 years earlier. But this seems to violate a basic norm of fairness. How could it have been fair for the court to apply the law retroactively to Chrysler, the manufacturer of the Henningsen's vehicle, and require the manufacturer to pay damages under a decision that was not announced until 5 years after the accident?[14]

Dworkin's account seems to provide a solution to this problem.[15] If the court was not really making new law, but only interpreting existing law, then it is a mistake to think that the court was applying the law retroactively. The court was just interpreting what the law had always been, not changing it.

But the impression that Dworkin's account solves the problem is an illusion. It is true that, if Dworkin's account is true, the unfairness involved cannot be described as a retroactive application of new law. But this does nothing to reduce the unfairness. If Dworkin's account is true, the unfairness just has to be described differently. Because on Dworkin's account, what had seemed to be a retroactive application of new law is not truly retroactive, I say that such cases involve *seeming* or *apparent* retroactive application of newly announced law, where I mean to leave it open whether the application of the newly announced law is truly retroactive.

Consider a case in which an appellate court overrules a well-established precedent on the basis of sophisticated legal reasoning. Suppose that the underlying cause of action arose out of an interaction between the plaintiff

and the defendant that took place years before the appellate decision. Suppose, also, that throughout the earlier interaction with the plaintiff, the defendant can show that it relied on the court's prior precedents and that it complied with the law as stated in those precedents. How can it be fair for an appellate court to overrule those prior cases years later and apply the newly announced law to the parties in the lawsuit in which the law is newly announced, thus requiring the defendant who faithfully complied with the law as enunciated at the time to pay damages to the plaintiff?

Dworkin himself suggests that decisions in these sorts of hard cases demand Herculean judicial skill. No one could think that it was reasonable to expect the nonlawyer parties to the lawsuit to have the skill to be able to reliably predict that an appellate court would overrule the prior precedents. Surely, nonlawyers have a duty to rely on and comply with established precedents until such time as they are overruled. Then how can it be fair for a judge to order that a nonlawyer defendant who has complied with and relied on established precedents must pay damages for failing to act in accordance with a principle or rule that was not announced until years later? Even if Dworkin is correct about what the law *is*, that does not make it fair to hold nonlawyer defendants to what the law is before they have any way of figuring out what it is.

To see how this seeming retroactivity of the civil law would be evaluated under the main principle, consider the alternative practice of not giving appellate court decisions of this kind apparent retroactive effect, but applying them only to causes of action arising after the decision has been announced. If this were the practice in civil appeals, no party to a civil suit would ever file an appeal if their appeal depended on the court's overruling prior precedent, because, even if the appeal were successful, there would be no benefit to the prevailing party. If no such appeals were ever filed, appellate courts would rarely have an opportunity to improve the law by overruling prior precedents. So if appellate courts do tend to improve the law when they overrule prior precedents (which I here assume to be true), and if the only relevant alternative is not to give such decisions apparent retroactive effect by not applying them to the case in which the new decision is announced, the main principle would favor the former alternative, which is the actual practice in civil law.

Note that this result can easily be explained by appeal to the expanded original position. Behind the veil of ignorance, everyone would recognize that they stand to benefit from the practice of giving apparent retroactive application of appellate decisions to the case between the parties to the appeal, because everyone would tend to benefit from the improvements in the law that would be generated by the practice, even those who would be unfortunate enough to be parties to an appeal in which the apparent retroactive application of the court's decision adversely affected them. Because the practice would only apply to civil suits, even the unfortunate defendants would only be liable for monetary losses, not imprisonment. Alternatively, if

appellate decisions in civil cases were not given apparent retroactive effect to the parties to the appeal, appellate courts would rarely have an opportunity to overrule past precedents and the process by which appellate courts improve the civil law would slow dramatically. This might significantly diminish everyone's life prospects. So the main principle can explain why such an apparently unfair practice is superior to the more "fair" alternative. I augment this discussion when I take up retroactivity in the criminal law in the next chapter.

Conclusion

We now have the tools for understanding moral progress and, by extension, legal progress. The understanding comes at the meta-level. There moral progress can be explained by exceptions to existing moral practices that satisfy the main principle's two-pronged consequentialist evaluation: where the exception is evaluated on its contribution to equitably promoting life prospects, when considered as both a substantive practice and a practice of implementation.

As a consequence of the paradox of direct consequentialism, the main principle does not endorse itself as a ground-level moral principle. Because the main principle evaluates practices in terms of a probabilistic measure of life prospects, ground-level moral norms and exceptions to them have a reference class logic. This makes it implausible to think that the process of moral improvement will ever end. There will always be a potential for further improvement.

The final element in the explanation of moral improvement is the development of individuals with what Aristotle called *practical wisdom*, which I understand as implicit sensitivity to the main principle. When a society or culture's moral practices produce at least some individuals with practical wisdom, I say that it has crossed the *consequentialist threshold*. Once a society or culture has crossed the consequentialist threshold, then changes to its moral practices are no longer random, from a moral point of view, because individuals who are implicitly sensitive to the main principle are able to discern at least some ways of improving the existing moral practices. I have no way of determining precisely when a society or culture crosses the consequentialist threshold, but there is a positive test for determining whether a society or culture has crossed it. If a society or culture's moral practices include a version of the Golden Rule, then the society or culture almost certainly has crossed it. Because all major religions and almost all others do accept a version of the Golden Rule, I am confident that all, or almost all, existing societies and cultures have crossed the consequentialist threshold. These societies and cultures are able to produce individuals with practical wisdom and thus, have the capability of improving themselves in a nonrandom way.

In Talbott (2005), I pointed to a list of basic human rights as an example of moral progress and provided a partial explanation of why legal guarantees of those rights are such an important moral improvement. In coming chapters, I extend that explanation in two directions. First, I deepen the explanation of the moral appropriateness of the basic human rights, by explaining more fully why they are supported by the main principle. Second, I introduce new human rights that, though not basic, are endorsed by the main principle and thus should be universally guaranteed to all human beings.

SIX

Security Rights

The first category of human rights to be discussed is, in one sense, the most essential. Security rights are on everyone's list of human rights. Moreover, it seems that there is a simple consequentialist case for security rights. Genocide, torture, murder, rape, and so forth are very bad for life prospects. Therefore, if a moral or legal code did not include prohibitions on these actions, the main principle would endorse changing it to do so. So, at first blush, the case for security rights seems trivial. Because, in addition, security rights will be endorsed by *any* remotely plausible account of human rights, it would seem that there is no point in spending much time talking about them.

However, this is a mistake. It turns out that when we look closely at the contours of these rights, they provide substantial support for a consequentialist account over a nonconsequentialist account. In this chapter, I continue the expository strategy of comparing the security rights endorsed by the main principle with those endorsed by libertarian natural rights to illustrate how the consequentialist main principle largely absorbs or supersedes whatever natural rights there may be, by endorsing exceptions to natural rights whenever the exceptions, evaluated as a substantive practice and as a practice of implementation, would do a better job of equitably promoting life prospects.

Because my conception of human rights is of rights that all governments should guarantee as robust and inalienable, my discussion will focus on the legal protection of rights. However, my discussion will indirectly implicate our moral beliefs, because our beliefs about what laws are and are not justified are moral beliefs.

I also continue to focus on primary ground-level norms, not on the secondary norms that deal with enforcement, except to the extent that the secondary norms affect the life prospects of those who cooperate with the primary ground-level norms. Thus, for example, I do not say anything about the appropriate punishment for those who violate security rights or the appropriate level of force permitted in the defending oneself against violations of one's security rights, but I do consider the potential for a legal enforcement practice to punish the innocent, whether intentionally or inadvertently, because that is a crucial element in the main principle's evaluation of a legal practice.

How Human Rights to Security Differ from Libertarian Rights

In this chapter, I continue my discussion of how the main principle can endorse exceptions to libertarian natural rights. A full analysis would require evaluating exceptions to libertarian rights as both substantive practices and as implementation practices. In this chapter, I continue my assumption that there is a legal authority that can change the laws in ways that do not generate a coordination problem, so that I can limit my discussion of the application of the main principle to the substantive evaluation of practices that I discuss.

It is useful to repeat the libertarian natural rights principle here:

Libertarian Natural Rights (With Enforcement Provision and Actual Consent Exception). Everyone has a natural right that others not intentionally or negligently cause them any basic harm (or the risk of a basic harm) and that others not threaten them with a basic harm (or the risk of a basic harm). When the relevant authorizing conditions are satisfied, everyone has a right to intentionally cause another person a basic harm (or the risk of a basic harm) or to threaten a basic harm as part of proportionate enforcement of a natural right—that is, in order to deter or prevent the violation of a natural right or in order to exact appropriate compensation or proportionate punishment for the violation of a natural right. Finally, any person may voluntarily waive or transfer a natural right, either conditionally or unconditionally.

In chapter 2 I specified the *basic harms* generally as personal harms (harms to one's body or mind) or harms to property. It will be useful to have a representative list of personal harms: being killed, tortured, physically assaulted, mutilated, disabled, raped, shackled, or imprisoned. Also included is being physically forced to do something against one's will or being subject to the psychological control of another (e.g., by brainwashing or hypnosis). The list is meant to be representative, not exhaustive.

From Libertarian Natural Rights to Human Rights

Libertarian natural rights are largely negative—they are rights not to be treated in certain ways by other people. The rights themselves include an enforcement provision that permits a person to cause or threaten basic harms to prevent a rights violation and to cause or threaten basic harms to exact compensation for and punishment of a rights violation. All of these factors operate to deter violations of libertarian natural rights. Call these deterrents the *built-in deterrents*, because they are part of the natural rights themselves.

There is no guarantee that these built-in deterrents will be successful deterrents. Indeed, we know that they will not be fully successful, because for all of human history, people have been inflicting basic harms on people

who have done nothing to deserve it. Most of the time, those on whom the harms were being inflicted tried to use force in self-defense or to punish those who inflicted harm on them, but that did not successfully deter their tormentors from harming them anyway. Following Locke, I refer to this failure of built-in deterrence as one of the potential *inconveniences* of the state of nature.

The shortcomings of the built-in deterrence in state-of-nature rights can easily be appreciated. Suppose we are in the state of nature. You threaten to punish me if I kill you. Such a threat would be an example of *individual deterrence*. Individual deterrence will be effective only if I don't think I can kill you, because you won't be able to punish me after you are dead. Call this the *failure of individual deterrence*.

Evolution seems to have developed a solution to the failure of individual deterrence, which I refer to as *kin deterrence*. In kin deterrence, you and the other members of your family threaten to punish anyone who kills you or any other member of your family. Kin deterrence of violations of natural rights against the intentional infliction of basic harms seems to be close to a cultural universal. Kin deterrence is a solution to the failure of individual deterrence, but it generates a new kind of problem. Because families have a potentially unlimited future, kin deterrence can generate potentially endless cycles of intentionally inflicted basic harms, illustrated by the example of the feud between the Hatfields and McCoys.[1]

Suppose a Hatfield kills a McCoy and claims to have acted in self-defense. If the McCoys don't accept the defense, then they will think that they must punish the killer. If the Hatfields do accept the defense, then they will think that they must punish any McCoy who kills the Hatfield who claims to have killed in self-defense. So if the McCoys don't accept the defense and the Hatfields do, the result could be a potentially endless cycle of tit-for-tat killings. Even worse, because all the McCoys are committed to punishing any Hatfield who performs a tit-for-tat killing and all the Hatfields are committed to punishing any of the McCoys who performs a tit-for-tat killing, *any* Hatfield who thinks the McCoys are contemplating a tit-for-tat killing could think himself justified in killing *any* McCoy in *self-defense*, and *vice versa*. This sort of situation can easily generate a collective action problem (CAP) in which the life prospects of the members of both families would be much higher if both sides were prevented from carrying out kin deterrence. I call this sort of CAP an *internal security CAP*. In this internal security CAP, a coercive policy that replaces kin deterrence with a criminal justice system could well promote the life prospects of all cooperators. When it did so equitably, the practice would be endorsed by the main principle, even if the families involved would not consent to it.

This simplified example can be modified to explain how groups larger than individual families (e.g., tribes, ethnic groups, nationalities, or religions) can provide similar solutions to problems of deterrence and how each such solution can generate potentially endless cycles of the intentional infliction

of basic harms. This is not a mere abstract possibility. These sorts of cycles can be observed throughout history and in many places today. But the simple example of kin deterrence contains all of the elements necessary to understand how the main principle has the potential to make natural rights irrelevant, so I focus on it.

The first point is that having natural rights against the intentional or negligent infliction of basic harms is nothing like a guarantee that one will avoid the intentional infliction of basic harms by others. Because basic harms are so bad, other things being equal, a system of coercion that effectively protects people from such harms will be endorsed by the main principle over a state of nature in which there is a substantial risk of suffering the intentional infliction of such harms. Hobbes used just such an argument in attempting to justify an absolute sovereign. I considered the shortcomings of his argument in the companion volume (Talbott 2005, chap. 7). Here I simply note that Hobbes's argument cannot be ruled out *a priori*. In a state of nature in which everyone's risk of suffering a basic harm was high (which was how Hobbes thought of the state of nature), any coercive practice that substantially reduced that risk would potentially benefit everyone ([1651], chap. 17). It is easy to see how such a practice could be endorsed by the main principle, if there were no superior alternatives.

Because the basic harms are so inimical to human well-being, other things being equal, the main principle will favor practices that reduce the probability of suffering basic harms. Moreover, because equity considerations give special weight to the less well off and those who suffer basic harms are *ipso facto* among the less well off, the main principle will give special weight to policies that reduce the probability of suffering basic harms. As a result, although it is not absolute, there is a strong presumption that any system of coercive laws endorsed by the main principle as an exception to libertarian natural rights will provide everyone with significant protection against the basic harms, if it is at all possible to do so.

In this way, the main principle goes beyond the libertarian natural rights against the intentional or negligent infliction of basic harms. The libertarian natural rights do not provide any guarantee of *protection* against the intentional or negligent infliction of basic harms; they provide only moral justification for attempting to protect oneself or others. But, as a practical matter, the main principle will *require* that everyone be assured some level of *protection* against basic harms, if it is at all possible to do so, for if some people are not protected against basic harms, the legal system will not equitably promote life prospects (cf. Shue 1980, 37–38).

My conclusion here is a practical one. If it really were impossible to provide any significant degree of protection against basic harms, then the main principle could not require it as a condition for its endorsement of a coercive legal system. But we know that, with few exceptions (e.g., temporary periods of great civil unrest), governments can protect their citizens against basic harms. So for a coercive legal system to be endorsed by the main principle, it must provide security against basic harms.

Why *Rights* to Security?

Why should the provision of security take the form of security *rights*? Why isn't government protection enough? A legal right implies some institutional mechanism for asserting the right. Why would the main principle discriminate between a legal right to security and other legal means of providing security?

There is no answer to this question *a priori*. The main principle evaluates practices on the basis of their consequences. Legal rights do a better job of protecting security. To see why, suppose that you are a ruler who has established a police force to protect your subjects. Some members of the police force are corrupt and are oppressing the people they are supposed to be protecting. Consider two possibilities:

(1) *No security rights*. There is no institutional mechanism by which those on whom basic harms are being inflicted by the police can claim to be entitled to relief. However, your subjects know you are a kind-hearted ruler. They write you a letter describing their situation. You empathize with their suffering and order the offending police officers to be removed from their jobs and punished. In this scenario, you are protecting your subjects from the intentional or negligent infliction of basic harms, but they have no right to the protection. It depends on your good will.

(2) *Security rights*. In this scenario, there is an institutional mechanism for subjects to use to complain when the police do not protect them from the intentional or negligent infliction of basic harms. The institutional mechanism need not be perfect. But it is necessary that it give them an institutional *entitlement* to protection against the intentional or negligent infliction of basic harms. For example, it might provide a legal entitlement to having the oppressive police removed from their jobs and punished, and perhaps a legal entitlement to the recovery of damages against the government for the police officers' failure to properly protect the citizens.

Hobbes thought that only an absolute ruler could effectively protect citizens against the intentional or negligent infliction of basic harms. We now have enough experience with other forms of government to know that he was mistaken. But it must be admitted that countries with governments that grant their citizens rights against the intentional or negligent infliction of basic harms sometimes have higher rates of violent crime between civilians than countries with governments that do not respect citizens' rights. Where rights seem to make the most difference is not in protecting civilians from basic harms inflicted by other civilians, but in protecting people from intentional or negligent *government* infliction of basic harms. The twentieth century has seen history's most spectacular examples of the intentional and negligent government infliction of basic harms—for example, Stalin's starving of millions of Kulaks and Hitler's extermination of millions of Jews, Roma, and Poles. It is this evidence of

the failure of governments to protect against their own intentional or negligent infliction of basic harms that provides the strongest grounds under the main principle for government guaranteed *rights* to protection against the intentional or negligent infliction of basic harms. It is almost inconceivable that Stalin could have killed millions of Kulaks or Hitler could have killed millions of Jews, Roma, and Poles if there had been legal institutions that entitled the victims to make a claim against the government to protection against the intentional or negligent infliction of basic harms.

For this reason alone, at least some rights should be universal—that is, rights to protection against the intentional or negligent infliction of basic harms. Such rights would be one part of a larger category of *security rights*. Security rights differ from libertarian natural rights against the intentional or negligent infliction of basic harms, because they are a right to *protection* against the intentional or negligent infliction of basic harms. In a state of nature, you may have a right that I not inflict basic harms on you and it may be *permissible* that others help you to enforce that right, but it is very implausible that in a state of nature you would have a *right* that other people help to protect you against violations of your rights. After all, protecting your rights could require them to risk their own lives. In the absence of some special agreement or other special circumstances, it does not seem that you could *require* others to risk their lives to help you enforce your rights. So security rights go beyond libertarian natural rights.

To guarantee security rights, a government must establish a criminal justice system to protect the innocent and to investigate, try, and punish rights violations.

Before the establishment of a government that protects security rights, each person in a libertarian state of nature would have a right to punish those who violated their rights; each person would have a right to defend herself against attempts to compel her testimony or to compel the production of evidence; and someone innocently accused would have a right to defend herself against attempts to compel her to stand trial. However, a criminal justice system endorsed by the main principle would include exceptions to the prohibition of coercion to allow for using coercion to compel testimony or the production of evidence or to confine suspects.[2]

Nozick (1974) famously argued that, in a state of nature, security rights could be provided by private protection agencies. These private protection agencies would eventually become a minimal state, the classical night watchman state. The main principle will endorse a state that goes well beyond Nozick's minimal state, because it will guarantee all the rights on my list of human rights. But even on the topic of security rights, the main principle endorses a state that differs significantly from Nozick's minimal state, because the security rights endorsed by the main principle will include a substantial number of procedural rights.

Procedural Rights

The design of a criminal justice system raises many difficult problems of institutional design. For example, because a government must assert and enforce a monopoly on coercive power, it is no easy task to design a system of government that will not abuse those powers. Abuses can be local—for example, excess brutality against marginal groups by police—or global—for example, arbitrary arrest and detention of opponents of the government. The potential to unknowingly convict and punish the innocent and the potential for abuse of the police powers of the state lead to a second category of rights that are a practical necessity for any system of security rights. These are *procedural rights*.

The first procedural right is the right to the status of a legal person before the law. This right might seem self-evident. Most other procedural rights are not self-evident. No one would ever claim that it is self-evident that people have a right to trial by jury or a right against self-incrimination or a right to the services of a defense attorney. None of these rights would obtain in a libertarian state of nature. Because they are generally not self-evident and because they exist only outside the state of nature, procedural rights provide a good model for understanding the rights endorsed by the main principle. *A priori*, one might have thought that confessions would be the most reliable method for determining guilt. It is due to actual experience—for example, the forced confessions in the Inquisition or the show trials under Stalin or during the Cultural Revolution in China—that we have discovered that judicial systems that rely on confessions (which, of course, they have the ability to force) tend to convict a high percentage of innocent people. The reason is simple. Most people subjected to torture confess. Thus, the fact that a person confessed under torture does not increase the probability that she is guilty. The right against self-incrimination is a protection against abuses that we have learned about from experience. Similarly, it is because individuals need protection against potential abuses of power by the state that defendants in serious crimes should have a right to the services of a defense attorney and a right to a trial by a jury of their peers.

A nonconsequentialist can agree with everything I have said, but insist that we don't need the main principle to explain why we should have protections against punishing the innocent. We should never punish the innocent. I compare how my consequentialist account compares with nonconsequentialist accounts on the issue of punishing the innocent shortly.

It would be difficult even to list all of the rights that comprise the category of procedural rights. They include the rights of those who are arrested or charged with crimes, the rights that fall under the category of rights to a fair trial, and the rights of prisoners.[3] Procedural rights raise explanatory problems not only for libertarian theories, but for most other nonconsequentialist theories, also. The reason is that they often don't fit with ground-level judgments of fairness. It is worth considering one example in detail, the practice

of granting limited retroactivity to court decisions in criminal cases that establish new procedural rights.

Retroactivity of Procedural Rights for Criminal Defendants

No reasonable consequentialist or nonconsequentialist theory would endorse a general policy of enforcing criminal laws retroactively. But because the content of procedural rights themselves is the product of a historical process of development, an important issue of retroactivity arises that illustrates the operation of the main principle.

The issue is illustrated by the famous case of *Miranda v. Arizona*.[4] In the *Miranda* case, the U.S. Supreme Court held that an effective right to non-self-incrimination required that police inform suspects in their custody of their right to remain silent, before interrogating them. This has come to be called the *Miranda rule*. Because the police who arrested the defendant Miranda did not comply with the *Miranda* rule (i.e., did not inform the defendant of his right to remain silent and to consult an attorney), the Supreme Court ruled his subsequent confession to be inadmissible, and on this basis, overturned his conviction for kidnapping and rape. But this seems very strange. There was no *Miranda* rule until the Supreme Court articulated it, which was long after Miranda's arrest and confession. It seems unfair for the Court to throw out his conviction on the grounds that the police did not comply with a rule that was not announced until after Miranda made his confession.

To understand why the court would make its decision retroactive to the Miranda arrest, we have to evaluate the various alternative practices. The first practice is the one of *full prospective application* of a new procedural rule: to apply the rule only to arrests made after the new rule is announced. I take it that it is obvious that there is a fairness argument for this practice. However, not everyone would agree that it was unfair to apply the *Miranda* rule retroactively to the Miranda case. Some would say that the Court was just articulating a rule of fairness that should have been complied with all along. Notice that this position implies that fairness requires a practice of *full retroactive application* of a new procedural rule: to apply the new rule to all past convictions, whether fully adjudicated or on appeal or not. I take it that it is obvious that this practice would have had a drastic effect on the criminal justice system. It would have required the Supreme Court to be willing to overturn almost all prior criminal convictions, because almost all of them would have been found deficient by the *Miranda* standard.

Now consider two additional practices: *qualified full prospective application* of a new procedural rule—to apply the new rule only to arrests made after the new rule is announced, but to also apply it to the case in which it is announced; and *limited retroactivity* of a new procedural rule—to apply a new rule only to cases not yet fully adjudicated or on direct appeal at the time the decision is made. It is hard to see how either of these two alternatives

could be defended on grounds of fairness. And yet it was one of these two rules that the Supreme Court followed in the *Miranda* case.

At the time of the *Miranda* decision, the Supreme Court's view was that the Constitution did not require retroactivity for new procedural rules; that retroactivity was to be decided on a case by case basis; but that if retroactivity were ordered, it should be full retroactivity.[5] The *Miranda* rule was not made fully retroactive.[6] It was given qualified full prospective application. In recent years, the Court has changed its position to hold that the Constitution does require some retroactivity for new procedural rules, but only *limited*, not *full* retroactivity.[7]

Let's consider the rule that was followed in *Miranda*, qualified full prospective application. Why would it be a superior rule, when evaluated under the main principle, than the fairer rule of full prospective application? To answer that question, recall the discussion of retroactivity in civil law in the previous chapter. In both the civil and criminal law, appellate courts will never get to make new precedents unless they can somehow motivate parties to appeal lower court decisions. For the Supreme Court to be able to announce new constitutional protections for criminal defendants, protections that potentially benefit lots of defendants, including many innocent ones, they need to motivate defendants to appeal constitutional issues to them. If new constitutional interpretations were given full prospective application, there would be no benefit to a criminal defendant to appeal a conviction, if the appeal would require overruling prior precedent. So there would not be any such appeals. Evaluated as a substantive practice under the main principle, full prospective application of new procedural rules would greatly impair the Supreme Court's ability to make improvements to constitutional jurisprudence and to much other jurisprudence besides. Nearly full prospective application is clearly superior.

What about the practice of limited retroactive application of new rules, which was subsequently adopted by the Supreme Court? First, in comparison with the practice of full retroactivity, it is obvious that limited retroactivity would be favored by the main principle, because full retroactivity would have the effect of releasing large numbers of convicts from prisons every time a new procedural rule was announced.

So the final comparison is between the two "unfair" rules: qualified prospective application and limited retroactive application. Qualified prospective application would generate a flurry of Supreme Court appeals of cases not yet fully adjudicated, because the only way that the defendants in those cases could get the benefit of the new rule would be by an appeal to the Supreme Court. This would place a substantial administrative burden on the Supreme Court. Giving their decisions limited retroactive application makes it possible for defendants to get relief in the trial court without the necessity of an appeal. This practice would produce results similar to the practice of nearly prospective application, but more efficiently. So the main principle would favor limited retroactive application. This is the rule that the Supreme Court has adopted.

What is important about the example is that the Supreme Court has used two rules, nearly full prospectivity and limited retroactivity, neither of which fits with ground-level fairness judgments, but both of which can be seen to be superior to the rules that do fit ground-level fairness judgments when evaluated under the main principle.

I should mention that it is open to nonconsequentialists to hold that the entire *Miranda* decision, not just the way that it was applied, was unjustified. They might hold that all new laws and all new interpretations of existing law should come from the legislature and thus be fully prospective and that courts should be limited to *passive* judicial review, simply determining whether the laws as written were followed. In a system of passive judicial review, there would never be any new judicial rules or new interpretations of existing laws, so the problem of retroactivity would not arise.

Of course, a consequentialist would be willing to compare a system of substantive judicial review with a system of passive judicial review. The question is this: Can the nonconsequentialist rule out active judicial review without considering the advantages and disadvantages of each? There is a myth that the U.S. Supreme Court is divided on this issue, but that is not correct. Conservatives on the Supreme Court are at least as likely as their liberal counterparts to invalidate legislation (Keck 2004). This is to be expected. Active review is almost certainly superior to passive review as a social practice of judicial review, because it is needed to avoid majority tyranny, as I explain in chapter 10.

Punishing the Guilty

It is easy to see how the main principle would justify many of the usual procedural rights, but there is one type of issue that typically causes problems for consequentialist accounts of justice: punishing the innocent. Before addressing this issue directly, let me consider a closely related issue that helps to introduce some of the relevant factors: punishing the guilty.

I have supposed that in the state of nature people have a right to punish those who have violated their rights. I have already mentioned that governments typically forbid vigilante justice, because the government asserts a monopoly on legitimate punishment. This seems unproblematic so long as the government takes responsibility for punishing those who violate a security right. How much punishment is the government justified in imposing? The main principle does not answer this question. I just assume that coercive threats and punishment satisfy a proportionality requirement, which is part of secondary rather than primary ground-level moral and legal practices.

Although the main principle does not provide an account of proportionality in punishment, it does make consequentialist considerations potentially relevant to questions of appropriate punishment. There is a striking contrast between the moral status of punishment in a retributivist theory like Kant's

and the moral status of punishment in a consequentialist theory like mine. Thus, for example, Kant believes that there is a duty to punish that admits of no exceptions.[8] But, at least in theory, the main principle does admit of exceptions. Suppose, for example, that it were discovered that psychotherapy was a much more effective method of reducing crime rates than punishment. In such a case, the main principle could endorse replacing all punishments with psychotherapy. To prevent vigilante justice, the government would have to *prohibit* punishing the guilty.

To see how the main principle could endorse prohibitions on punishing the guilty, we have to consider a possible world very different from the actual world, a world in which psychotherapy was effective in reducing the crime rate only if potential criminals were not threatened with punishment. For example, it might be that the threat of violence involved in legal sanctions would itself provoke violent acts that would not be provoked if there were no legal sanctions. In this world, one that is, admittedly, very far from the actual world, the main principle would endorse an exception that mandated universal psychotherapy for criminals and prohibited punishment, not for its beneficial effects on criminals, but for it benefits to potential crime victims. The life prospects of potential victims would be greatly increased if crime were largely eliminated.

In such a world, if punishment were replaced with psychotherapy, murders would occur only very rarely. Would murderers *deserve* to be punished? The main principle is silent on that question. But the main principle would endorse a practice that prohibited punishing them, even if they deserved it. This is an important difference from at least some nonconsequentialist theories.

Punishing the Innocent

There is a family of examples that raise problems for utilitarianism, at least direct utilitarianism. They are examples of the following kind. A serious crime has been committed. The community is upset and will be much happier if they believe that the perpetrator has been captured and punished. The sheriff has in custody a suspect whom he knows to be innocent, but he can plant evidence or otherwise cook the evidence to make the suspect appear guilty. Because the sheriff has no prospect of finding the actual perpetrator, he cooks the evidence, the innocent suspect is convicted, and the community is much happier. Because direct utilitarianism allows for the sum of the good feelings in the community to outweigh the harm to the innocent suspect, it gives the result that, if the community is large enough, the sheriff has a duty to plant evidence and obtain a conviction of the innocent suspect.

The main principle does not permit this kind of aggregation. It requires that a practice equitably promote the life prospects of everyone who cooperates with the relevant practice, including the innocent victims who would be framed. In the expanded original position, from an impartial point of view,

we would not balance the satisfaction of the community against the imprisonment of the innocent victim. The consequences of a mistaken conviction and imprisonment are bad enough that, because the main principle is a prioritarian principle that gives special weight to the life prospects of the less well off, the satisfaction of the community cannot outweigh the reduction in life prospects for the person falsely imprisoned. An additional consideration is that a system that permitted false convictions would surely be abused and produce lots of them.

A nonconsequentialist, however, could have a much simpler explanation of this result. The nonconsequentialist could insist that the act of intentionally convicting the innocent has infinite negative weight. This seems to me to be a mistake, as I explain shortly.

To fully understand how the main principle applies to this example, we must consider two ground-level practices. First, there is the ground-level legal practice of making an exception to criminal law to allow cooking the evidence to convict a defendant known to be innocent under circumstances in which there is a substantial social benefit to doing so. I want to be able to use the main principle to explain why such an exception would not be an improvement in the criminal law. Second, there is the ground-level moral practice of condemnation of the very idea of an evidence-cooking exception to criminal law. I want to be able to use the main principle to explain why changing that attitude would not be a moral improvement in ground-level moral practices.

Consider the attitude of condemnation first. When we look at the ground-level moral practice of condemnation of the cooked evidence exception from the inside, as it were, we find a strong condemnation of the idea of intentionally convicting a person known to be innocent, a condemnation so strong that it seems to be best explained by the idea that intentionally convicting the innocent has infinite negative value, and thus it is never to be done.

What we want to do is to step outside that moral practice of condemnation and ask this: What is the best explanation of it? One explanation, a nonconsequentialist explanation, takes the ground-level practice at face value and says that the explanation of the ground-level practice is that intentionally convicting an innocent person is infinitely bad. This is a potential explanation. Is it the best explanation?

There are puzzles for this view from two directions. First, consider two different cases of cooking the evidence to convict an innocent defendant. In the first, the defendant is charged with murder and the penalty is life imprisonment. In the second, the charge is breaking and entering and the penalty would be 6 months in jail.

In both cases we have a strong ground-level feeling of condemnation for intentionally convicting an innocent person. But it is hard to think that the explanation of that feeling is that both cases involve something of *infinite* negative value. It seems clear that if an attorney had time to file a successful appeal

in only one of the two cases, she should choose the case of the innocent murder defendant rather than the innocent breaking and entering defendant.

The nonconsequentialist can reply that it is not the result of convicting an innocent person that is of infinite negative value, but the act of intentionally cooking the evidence to bring about the conviction. But again, if you were coerced to cook the evidence on one case or the other, wouldn't it be better to cook the evidence on the breaking and entering case? This is one side of the puzzle for nonconsequentialists raised by the example of convicting the innocent.

Unintentionally Convicting the Innocent

The other side of the puzzle emerges when we ask about unintentionally convicting the innocent. Does the nonconsequentialist have the resources to explain how a legal system that unintentionally convicts some innocent defendants could be justified?

Let's begin by considering whether the main principle has the resources to be able to endorse a legal system that convicts some innocent defendants. The main principle's substantive evaluation of a practice is based on the extent to which it equitably promotes the life prospects of those who cooperate in the practice. Innocent defendants cooperate, so their life prospects are included in the evaluation. Consider the life prospects of someone who is falsely convicted of murder and spends most of her life in prison, generally thought to have been guilty of a murder. This is a pretty bleak prospect. Because the main principle gives extra weight to those who are less well off, there will be a strong, but not absolute, presumption against systems that convict innocent murder defendants.

How could the main principle endorse a system that made it inevitable that some innocent defendants would live out their lives in prison, believed by their friends and perhaps even their families to have committed a murder?

Consider how the question would be addressed in the expanded original position. Would we agree to a criminal justice system that we knew would convict some innocent defendants? In the expanded original position we would want to know: How many innocent defendants would be convicted? What would be the alternative? No criminal justice system? How many murders and other basic harms would be deterred or prevented by the criminal justice system that convicted some innocent defendants? Even if it would be awful to be wrongly convicted of and punished for a murder one didn't commit, being murdered is pretty bad, too. In the expanded original position, I think it is clear that it would be reasonable to agree to some nonnegligible risk of being wrongly convicted of murder in order to substantially reduce the risk of being murdered. In any case, it is clear what the decision would turn on in the expanded original position. It would be necessary to balance increases in the probability

of a truly awful life prospect (being mistakenly convicted) against decreases in the murder rate (and thus the probability of being murdered).

This issue is of more than mere academic interest. Our own legal system in the United States is one in which aggressive police and prosecutorial practices can bias the evidence enough to obtain a conviction. We know that this leads to the conviction of innocent defendants, even though the police and prosecutors typically don't intend to convict an innocent person. In the past, discussions of wrongful conviction have taken place in a vacuum, because there was no way to estimate what the rate of wrongful conviction was. With the introduction of DNA evidence, for the first time, we have an independent test for wrongful convictions for crimes for which DNA evidence is available. So far, the test has primarily been employed in capital crimes.

In the most careful statistical study thus far, Risinger estimated that the minimum rate of wrongful conviction for defendants in capital rape-murder cases in the period from 1982 to 1989 was 3.3%—approximately 1 in 30 defendants convicted—and the maximum rate was around 5%, or 1 in 20.[9] Rape-murder cases were selected because they are the cases in which DNA evidence is most likely to be available, so they provided the largest sample. Is 3.3% too high? Is 5% too high?[10] A consequentialist would need to know whether it could be reduced without significantly increasing the murder rate. How would a nonconsequentialist decide?

Or consider another question. In many criminal cases, the victim knew the perpetrator or there was a voluntary confession containing details that could not have been known except to the perpetrator or there was other definitive evidence of guilt. Presumably, the rate of wrongful conviction is much lower than 3.3% in those cases. Set them aside. Consider all the cases in which the evidence is not definitive, including convictions based on eyewitness identifications of strangers, "confessions" that contain no corroborating information or circumstantial evidence. The rate of wrongful conviction in these cases must be higher than 3.3%. It is almost surely higher than 5%.[11] Could anyone think that convictions in such cases were beyond reasonable doubt? What should the standard of proof be for criminal trials? A 5% wrongful conviction rate? Lower? Higher? For a consequentialist, the answer will depend on how effective the system is at deterring murder. How could a nonconsequentialist answer these questions?

Nonconsequentialists typically have an easy time of explaining why no system that intentionally punishes the innocent can be justified, but what can they say about systems that do so unintentionally? It would not be plausible to require that we minimize the probability of a wrongful conviction, because the alternative that minimizes the probability of a wrongful conviction is to have no criminal justice system at all.

Would the nonconsequentialist hold that "Better that 10 guilty men go free than that one innocent person be convicted"? This is not even a relevant consideration. The relevant information is information on how many potential murders would be deterred.

Echoing Kamm (1989), Nagel (1991, 148) suggests that we should be willing to accept an increase in the murder rate for a criminal justice system that reinforces our moral status as rights holders. This consideration is one that might be accommodated by the main principle. After all, if reinforcing our moral status as rights holders reduces the sorts of abuses that lead to conviction of the innocent, then, other things being equal, the main principle would endorse reinforcing that status. But Nagel and Kamm seem to think of moral status as an independent consideration, to be added in after all the consequentialist considerations have already been taken into account. This puzzles me. Suppose that reinforcing our moral status costs only one additional murder. How would Nagel and Kamm explain their rationale to the family of the person murdered? Would they say that the importance of reinforcing the moral status of everyone who wasn't murdered outweighed their loss and the loss of the murdered person? This seems to be a case in which their theory would have the same sort of aggregation problem as utilitarianism.

Would they say that moral status has a value incommensurable with the value of a murdered loved one? How could it be that valuable? Don't misunderstand me. I think that the status of being regarded as fully human is immensely valuable. Look at all the awful things that have been done to people who were regarded as less than fully human. Even when there is no physical harm, think of the psychological harm from being regarded as less than fully human. However, all of these kinds of harms make a difference to well-being, so they are all taken into account by the main principle. The question is whether after all the losses to well-being have been considered, there is some other loss that has been left out. What could it be that would justify failing to prevent a murder?

The consequentialist insists that any adequate account of this example will require some way of evaluating alternative practices that involves balancing increases in the probability of being mistakenly convicted against decreases in the crime rate. If the nonconsequentialist prohibits balancing, it is hard to see how the view could address cases of this kind. If the nonconsequentialist permits balancing, what resources can there be to explain how to do it?[12]

The Doctrine of Double Effect

One nonconsequentialist doctrine that is often appealed to in cases such as this is the doctrine of double effect (DDE). The basic idea of DDE is that it is permissible to cause harm if it is a foreseeable side effect but not if it is a goal or the means to a goal.[13] Of course, not just any bad side effects can be justified under DDE. There is a proportionality requirement—that is, a requirement that the intended good be proportionately greater than the merely foreseen but unintended harm. Thus, according to DDE, it would never be permissible to intentionally kill an innocent person, but it may be permissible to cause the unintended

but foreseeable deaths of innocent people as part of a criminal justice system that is aimed at achieving and in fact achieves proportionately greater good.

To begin with, it is important to see that DDE or something like it is exactly the kind of ground-level moral principle that we would expect to be favored by the main principle, with one qualification that I mention shortly. There is overwhelming historical evidence, including evidence from the Marxist dictatorships of the twentieth century, that permitting human beings to intend *bad ends* in order to promote what they judge to be proportionately better ends, or to intend *bad means* to what they judge to be good ends can be disastrous. So the main principle would support ground-level moral principles that create a strong presumption against intending bad ends or bad means to good ends. Also, it seems clear that the proportionality requirement introduces an explicitly consequentialist element into DDE. So some reasonable, defeasible DDE doctrine would be explained by the main principle.

Nonconsequentialists typically regard DDE as absolute or, in my terms, categorical—that is, it admits of no exceptions.[14] For the reasons discussed in chapter 5, we would expect that the main principle would endorse some exceptions to DDE. It is not only consequentialists who allow exceptions to DDE. Nozick's natural rights theory allows killing in self-defense not only of those who intend to kill you, but of innocent shields who, against their will, have been placed so that you cannot stop the person intending to kill you without killing them. In one example, Nozick even allows that you may vaporize someone to prevent them from falling on you, even if someone else pushed them to cause their fall (1974, 34–35). Clearly, this would violate DDE.

Another kind of exception that would surely be acceptable to any nonconsequentialists who use an original position test would be what I refer to as *Sophie's choice exceptions* (Styron 1979). Williams (1973, 93–99) provides an example, which I modify. A group of terrorists have taken ten innocent hostages, a husband and wife and eight children, and intend to kill them all. They have already killed two of the children when you arrive on the scene. To honor you, they offer to free the other seven hostages if you will kill one. They give you a pistol with only one bullet in it, so there is no way for you to kill all the terrorists. The father pleads with you to kill him to save the rest of his family. Would it be permissible for you to kill him?

Surely a narrowly drawn Sophie's choice exception would be unanimously agreed to in any plausible original position. The only question would be whether the application of the exception would lead to unintended and undesirable consequences. However, the Sophie's choice exception would not invite abuse. Also, evaluated as an implementation practice, unilaterally acting on the exception would not generate a procedural coordination problem. Therefore, it seems clear that this exception would be endorsed by the main principle. So I conclude that the main principle explains DDE, if it is understood as a defeasible principle. If it is understood as absolute or categorical, it has implications that we should reject.

More on Intentionally Convicting the Innocent

The discussion of unintentionally convicting the innocent was just one branch of my discussion of the problem of convicting the innocent. It was to raise a puzzle about how nonconsequentialists are to evaluate the badness of an erroneous conviction. It seems to me that the consequentialist has the only plausible way of thinking about that question.

But if the main principle allows for convicting the innocent to be less than an infinite bad, doesn't that imply that it will endorse a practice in which police and prosecutors occasionally intentionally cook the evidence against an innocent person, if not to relieve anxiety, then at least to prevent a large number of other murders? Undoubtedly, there are some possible circumstances in which the main principle would endorse a practice of intentionally convicting or even intentionally killing an innocent person. That follows from my insistence that there are almost always exceptions to any moral norm. The Sophie's choice examples illustrate one example of this kind.

But I think that the main principle can also explain the appropriateness of the ground-level moral attitude that would make us react with shock and outrage to the suggestion that police or prosecutors should be able to make an exception to the norm against intentionally convicting an innocent person. For the same reason that giving governments the power to ignore individual rights and directly apply the main principle has led to disastrous results, giving police and prosecutors the green light to use their judgment to decide when to intentionally convict an innocent defendant would lead to great abuse.[15] I believe that that attitude of shock and outrage, which feels like an assignment of an infinite negative value to intentionally convicting the innocent, is instead a morally appropriate reaction to the potential for abuse of police and prosecutorial power, a reaction that would be readily endorsed by the main principle.

Procedurally Determined Liability and Strict Liability

We usually think of our legal system as one that bases liability on fault. But this is not quite correct. A criminal justice system defines a certain procedure for determining guilt or innocence and then imposes liability on the basis of the results of that procedure. If the procedure always convicted guilty defendants, then it would be accurate to say that the system bases liability on fault. But because any actual criminal justice system will convict some innocent defendants, it is more accurate to say that a criminal justice system bases liability not on fault but on the results of a procedure for determining fault. Let us say that systems that base liability on the results of a certain procedure are systems of *procedurally determined liability*. The question we are considering is when systems of procedurally determined liability are favored by the main principle.

When we ask the question this way, we are confronted with what seems like a potential *reductio* of the main principle. The main principle distinguishes between responsible noncompliers and nonresponsible noncompliers, but it does not require systems of procedurally determined liability to do so. Is it possible that the main principle could endorse a system of procedurally determined liability that made no procedural determinations of fault? Could the main principle endorse a system of strict criminal liability? If so, this would seem to be a *reductio* of the main principle. Surely, it would seem, no acceptable theory can endorse a system of strict criminal liability.

In the next section, I show that our suspicion is correct and that the main principle could endorse a system of strict criminal liability. I then reverse the argument and try to show that in the circumstances in which such a system would be endorsed by the main principle, it really would be an improvement over a system based on fault. If my argument is successful, it will raise a challenge to almost all nonconsequentialist accounts, because almost all of them rule out from the beginning the possibility of justifying a criminal justice system not based on fault.

How the Main Principle Could Have Endorsed a System of Strict Criminal Liability

We don't have to imagine a far-fetched possible world to understand how a system of strict criminal liability could be endorsed by the main principle. Surprisingly, a good case can be made that, in circumstances in which it would be endorsed by the main principle, such a system really would be a moral improvement and might well be morally superior to a legal system based on fault.

Consider, for example, Sophocles' play *Oedipus the King*. At the beginning of the play there is great sorrow in Thebes, which is being punished by the gods for not having avenged the death of Oedipus's father Laius. At the end of the play, when Oedipus discovers that he unknowingly killed his father and married his mother, he punishes himself by blinding himself. This makes for powerful theater, but to our contemporary moral sensibilities, it is puzzling. We must wonder why Oedipus was so hard on himself when it was his father who struck him first. Furthermore, when he retaliated he didn't know that it was his father he was striking. Nor did he know that Jocasta was his mother when he married her. And remember that his parents had sent him away to be killed as a young baby, so it can hardly be said that he owed them a debt of gratitude. In the play, all the characters seem to assume that Oedipus is subject to a standard of strict liability for killing his father and marrying his mother.

I hope that you can see that my discussion of the Oedipus story appeals to many kinds of exceptions to norms that we take for granted, but that may not

have been so obvious to the ancient Greeks. There is another example of strict liability in the Oedipus story. When, at the beginning of the play, Oedipus places a curse on the person who killed his father, the audience knows that he is placing a curse on himself. But we have to wonder if there shouldn't be an exception for curses when the person making the curse does not realize he is the person being cursed.

Similar stories can be found in many different moral traditions. To take one more example, consider Saul's curse of his son Jonathan in the Old Testament. On the day of battle, Saul had cursed anyone who tasted of food before evening. Jonathan had no knowledge of the curse, so he ate some honey. When he found out about the curse, both Saul and Jonathan assumed that he deserved to die. Jonathan was saved only by a great public outcry (I Samuel 14: 24–45). As I see it, the public outcry represents the possibility of a democratic process for making exceptions to traditional norms, a practice that would be endorsed by the main principle. But the norm itself is a norm of strict liability.

One further example of strict liability can be found in the norms that require suicide to protect honor in many moral traditions. Typically, as in hara-kiri in traditional Japan, this is based on strict liability without fault. Again, it seems plausible that these norms could be improved by incorporating exceptions for lack of fault into them.

These examples indicate to me that even the concept of liability based on fault may itself be a moral development that resulted from exceptions to earlier moral norms of strict liability without fault. What is most important about the examples is that they are examples of norms of strict liability that may well have been endorsed by the main principle at the time they were first established, because they may well have increased the life prospects of cooperators, both compliers and nonresponsible noncompliers. Even more surprisingly, it is a contingent matter whether replacing these norms of strict liability with norms of liability based on fault would have been an improvement under the main principle. Let me explain.

If the status quo is one in which there is no system of criminal justice, a system of strict criminal liability could easily be an improvement under the main principle. Of course, some nonresponsible noncompliers would be punished under the system. But, as we have already discussed, in any realistic world, some *compliers* will be punished by any criminal justice system. If a system that punishes some compliers can be justified, then there is no reason in principle why a system that punishes some nonresponsible noncompliers could not be justified.

What would seem to prevent a system of strict criminal liability from being endorsed by the main principle is that the main principle requires that a change not only be an improvement over the status quo, it must also be superior to the relevant alternatives. If one of the relevant alternatives is a system of liability based on fault, could the main principle ever endorse a system of strict criminal liability?

The answer is yes. To see why, consider a very simple system of criminal liability. There is only one prohibition, a prohibition of murder, and there is only one kind of nonresponsible noncompliance, killings in fits of insanity. Consider two criminal justice systems, one that allows an insanity defense and one that imposes strict liability for killing another person with no insanity defense. To keep the example simple, let's suppose that neither system ever punishes someone who has not killed. Any nonconsequentialist principle of justice based on fault would have to exclude the system that did not allow an insanity defense. But the main principle could favor the system of strict liability.

Suppose, for example, that murderers were so good at mimicking the symptoms of insanity that there was no reliable way to distinguish the truly insane from the impostors. The system that allowed the insanity defense might let off enough murderers that it would not effectively deter murder and the murder rate would be much higher than under the system of strict liability with no insanity defense. If the difference were great enough, the life prospects of all cooperators, composed of compliers and nonresponsible noncompliers (i.e., those who actually kill someone in a fit of insanity), might well be higher under the system of strict liability with no insanity defense than under the system of liability based on fault with an insanity defense. In such a case, the main principle would endorse the system of strict liability over the system of liability based on fault. Isn't that the correct ranking?

We can easily generalize this simple example. If determinations of fault were unreliable enough, a system of strict criminal liability might be favored by the main principle over a system of liability based on fault. The determination would turn on exactly the same kinds of considerations that determine whether the main principle will endorse a criminal justice system that inadvertently punishes some innocent defendants.

So now we have a way of understanding the historical development of systems of strict criminal liability. They might well have been endorsed as a moral improvement by the main principle at the time that they developed. In addition, we can see that a system of liability based on fault would not have been an improvement over a system of strict criminal liability until there were reliable procedures for determining fault. If no reliable procedures for determining fault had ever been developed, we might still have a system of strict criminal liability today and not think it at all unjust.[16]

Because criminal penalties such as imprisonment are not involved, it is much easier for the main principle to endorse strict liability in civil law. Over the past 200 years and especially the past 50 years, there has been a revival of strict liability in the civil law. I discuss this development in chapter 9.

Organ Harvesting

There is another kind of case that has been thought to raise problems for utilitarianism. These are the organ harvesting examples, in which, for example, five

lives can be saved by killing one healthy person and distributing his organs to five who would die without them. This kind of example does not raise problems for my account, because such a practice would not improve the life prospects of everyone, because it would *reduce* the life prospects of the organ donors.

There is more to be said about this sort of case, however, because it also can be used to illustrate the advantages of indirect consequentialism over direct consequentialism. The direct consequentialist would have a powerful reason for compelling the organ donation, if five socially useful lives of normal length would be saved and only one would be lost. But a social practice consequentialist could never justify such a practice, at least not in any world remotely resembling the actual world.

For one thing, the potential for abuse would be huge. But even if there were no danger of abuse, evaluated as a practice, organ harvesting would have disastrous indirect consequences. Thomson mentions that the practice of organ harvesting would tend to make people less careful about caring for their health (1990, 184). This is true, but it greatly underestimates the problems that such a practice would generate.

Suppose, for example, that the policy is to always provide those whose organs are failing with organs from the healthiest "donors" in their 20s. This policy would provide a powerful incentive to be less healthy than the norm among one's peers by age 20. Under such a practice, it is easy to imagine that the life expectancy of those less healthy than the norm in their 20s would be *higher* than the life expectancy of those more healthy than the norm. Thus, there would be a great incentive to impair one's health before age 20.

But the problem is worse than this. The more young people there were who attempted to make themselves less healthy than the norm, the more the norm for health would decline. If organs were supplied to all who needed them, there would be a great incentive for young people to damage their organs, thus assuring that they would be an organ recipient and not a donor. Thus, the practice of organ harvesting would produce a downward spiral in health, as the competition among young people to make themselves less healthy than the norm caused progressive declines in the norm itself.

Such a decline in the norms for health would reduce, not promote, people's life prospects. So no reasonable form of indirect consequentialism would favor a social practice of organ harvesting.

Conclusion on Security Rights

Previously, I explained why the main principle would endorse security rights that go beyond libertarian natural rights against the intentional or negligent infliction of basic harms. These security rights would go beyond the corresponding libertarian natural rights in three ways: (1) they would supersede some parts of libertarian natural rights, for example, forbidding vigilante justice; (2) they would include institutional guarantees of enforcement, adjudication, and

punishment, so they would require a police force, a judicial system, and a penal system; and (3) they would themselves require institutional guarantees of further procedural rights. Security rights should include protections not only against the intentional or negligent infliction of a basic harm but also protections against the intentional or negligent imposition of the risk of a basic harm.

Thomson's Exceptions to Natural Rights

It is useful to compare my theory of exceptions to natural rights with Thomson's (1990). Thomson allows four kinds of exceptions to natural claims and other kinds of natural rights. Two of them are giving one's word and promising (1990, 348–352). She refers to these as *word-giving* and *consent*. However, because giving one's word is a voluntary act by which one gives another person a claim, I categorize it as a consent exception also. So both of these exceptions fall under the actual consent exception introduced in chapter 2.

On Thomson's account, there is another source of nonnatural claims as well as a source of exceptions to natural rights: the claims that arise from legitimate laws (1990, 354–358). Thomson thinks it is implausible to think that all legitimate laws arise from actual consent of the governed, and she thinks that hypothetical consent fails to explain "whatever it is about a thing that makes it worthy of consent by the person" (360). Unfortunately, she does not provide any explanation of what that is either. So she provides no way of distinguishing between legitimate and illegitimate government action. She simply notes the exception.[17] I return to this exception shortly.

The fourth kind of exception that Thomson allows for is a consequentialist (though not utilitarian) one, given by her revised trade-off idea. Abstracting away from complications that will not be relevant here, the most important feature of the revised trade-off idea is that it permits exceptions to natural claims when an infringement would produce a "sufficiently large and appropriately distributed increment of good, or advantage" (1990, 197).

In its appeal to increment of good or advantage, appropriately distributed, it resembles the main principle, which permits exceptions to natural rights for practices that equitably promote the life prospects of all those in the relevant group. *Advantage* is Thomson's term for estimates of well-being based on expectations, rather than actual outcomes (1990, 170). She includes *advantage* in the revised trade-off idea for the same reason that I include *life prospects* in the main principle. No principle stated in terms of actual well-being could ever be applied, because we are never in a position to be able to guarantee an actual level of well-being, only better or worse prospects for well-being, or in Thomson's terms, advantage and disadvantage. In the remainder of this section, I explain the most important reason that the main principle is an improvement on Thomson's revised trade-off idea. Then I explain how it subsumes her other three exceptions, also.

First, a problem about Thomson's application of the revised trade-off idea. The main principle applies to social practices. Thomson does not say exactly what the revised trade-off idea applies to. She implicitly interprets it indirectly, as applying to practices, rather than to acts, without ever addressing the question. So I begin by explaining why the revised trade-off idea must be applied to practices rather than acts.

Thomson illustrates the revised trade-off idea with a variation on the well-known trolley example:

> An out-of-control trolley is hurtling down a track. Straight ahead of it on the track are five men who will be killed if the trolley reaches them. Bloggs is a passerby, who happens at the moment to be standing by the track next to the switch; he can throw the switch, thereby turning the trolley onto a spur of track on the right. There is one man on the spur of track on the right; that man will be killed if Bloggs turns the trolley. (1990, 176)

On Thomson's view, because throwing the switch will cause the death of the one person on the spur line, it is a serious infringement of that person's natural claims. Ordinarily, infringing a person's rights is morally impermissible. For Thomson, the revised trade-off idea defines a narrow exception in which infringing a natural right is *permissible*. (She does not think there are any circumstances in which Bloggs is morally required to throw the switch, because he may have conscientious objections to doing so [1990, 196].)

Thomson presents one version of the trolley example in which she claims it would be permissible for Bloggs to throw the switch. The example is one in which all six workers had been assigned to their positions by lot. Call it the *assignment by lot version* of trolley. In this version, Thomson claims that what is crucial is that there was a time (before work assignments had been made, when each had an equal chance of being the lone worker on the spur or any one of the five on the main track) at which the workmen all would have consented to the act of Bloggs's turning the switch; and that the morally relevant feature of the situation is not their hypothetical consent per se but the fact that, at that earlier time, it was to the *advantage* of each that Bloggs turn the switch—to be exact, that there is a sufficiently large and appropriately distributed increment of advantage sufficient to justify Bloggs turning the switch (182 and 192).

I agree with Thomson that it is permissible for Bloggs to turn the switch in this version of the example. But I do not believe she has adequately explained why. To see the problem with her explanation, I need to specify more details of the example. At 3:00 PM, when Bloggs has to decide whether to throw the switch, the trolley is hurtling toward Art, Bob, Carl, David, and Earl, while Fred is alone on the spur line. What exactly is Bloggs doing? He is throwing the switch to divert the trolley from the line that Art, Bob, Carl, David, and Earl are standing on and to divert it to the line that Fred is standing on.

Earlier in the morning, before the job assignments were made, that action, so described, was not to the advantage of Fred. It would be to his great disadvantage. So that act does not pass Thomson's test. It is not to the advantage of all the workers at any time.

Of course, Thomson was not thinking of the act described in that way. She was thinking of what I have called the *practice* of throwing the switch to save five and kill only one. It is true that early in the morning, before work assignments are made, that practice is to the advantage of everyone. But then we must ask, why is that the relevant practice to be evaluated, the practice of throwing the switch to save five and kill only one, rather than the practice of killing Fred to save Art, Bob, Carl, David, and Earl? This is a manifestation of what I have referred to as the problem of descriptive relativity. I think she needs to take advantage of the analysis in terms of social practices to solve that problem. In any case, that is not the most serious problem with Thomson's account.

Let's restate the lesson Thomson draws from the assignment by lot version of the trolley example: If there is a time *t* at which practice P would produce a sufficiently large and appropriately distributed good or advantage for the members of a group G, then it is permissible to act in accordance with practice P even though it infringes natural claims of some members of the group (199–200). Call this *Thomson's hypothetical consent exception to natural claims*. This exception is stated generally, but we can see from its structure alone that there is a problem with it. The problem is due to the time requirement. Because the exception only requires that there be a single time when the practice P would produce a sufficiently large and appropriately distributed good or advantage (understood as an expectation) for the members of group G, it is a single-time-slice principle. As the discussion of the maximin expectation formula and other principles with cutoff dates (or times) in chapter 4 illustrated, no such principle can be adequate.

Recall the example from chapter 4, in which the social practices gave members of the most advantaged group (MAG) high expectations at birth. I argued there that those social practices might not be justified if, though all had high expectations at birth, after birth some of the members of the MAG were likely to do much better than their expectations at birth and others were likely to do much worse. Any adequate principle would have to pay attention to more than a single time slice. Because Thomson's hypothetical consent exception to natural claims is a single-time-slice exception, it would give the wrong verdict on such examples.

Consider the following addition to Thomson's own example. Suppose that work assignments had been made by lot at 8 AM. Then at 10 AM management decided that those work assignments would be permanent—that Fred now has a permanent assignment to work alone on the spur line. Does the fact that the practice of throwing the switch was to Fred's advantage at 8 AM make it permissible for Bloggs to throw the switch at 3 PM? Or does the fact that the practice was not to Fred's advantage at 10 AM make it impermissible for

Bloggs to throw the switch? I see no way that a single-time-slice principle can solve these problems.

Thomson's single-time-slice principle seems to me to be an imperfect attempt to capture the idea that throwing the switch is permissible if doing so is part of a practice that equitably promotes the life prospects of everyone involved. If one kind of worker were always sacrificed to save others, the practice would not equitably promote the life prospects of everyone involved. The problem is that Thomson's single-time-slice principle is not an adequate test. As we have seen, the main principle evaluates life prospects at every stage of life.

What rule would the main principle endorse for cases like this? Even to ask the question in this way is to make it impossible to see how the main principle would apply.

For decades, philosophers have attempted to formulate exceptionless ground-level moral rules that would permit Bloggs to throw the switch in this example, but would not permit Bloggs to push a large man in front of the runaway trolley to stop it and would not permit Bloggs to harvest organs from a healthy person to save five potential recipients, as well as applying to all the variations on variations of these examples in the literature. It is hopeless, because, as I explained in chapter 5, there are almost always exceptions to any moral norm or principle endorsed by the main principle. Exceptions are an expected consequence of the reference class logic of moral norms and principles.

However, there is another reason that the search for exceptionless ground-level norms and principles is doomed to fail in this kind of case. The main principle evaluates moral practices rather than rules, because, good moral practices don't just produce good rule followers, they produce people with moral sensitivity or what Aristotle called practical wisdom. As I explained in chapter 5, moral sensitivity is a kind of implicit sensitivity that cannot be duplicated by explicitly applying rules. It can no more be replaced by a rule book than could a good center fielder in baseball be replaced by a physicist who knows the laws of physics that govern projectile interception (catching fly balls). Moral rules are helpful for learning moral sensitivity, and they play an important role in solving coordination problems, but they are only a small part of any moral practice.

When we respond to trolley examples or organ harvesting examples, we are exhibiting the responsiveness that we have developed from our moral training. That responsiveness often cannot be captured in a ground-level rule that one could use to make decisions. As discussed in chapter 5, a good moral practice will "silence" many morally discreditable reasons. Thus, it is an important feature of our moral practice that it "silences" the thought that another person might be used to block a runaway trolley. To see the great advantages of our practice, just imagine that we had a different moral practice in which people were told that, though it is very rarely justified, we should always be on the lookout for opportunities to use other

people as means to prevent worse outcomes. Not only would there be much more lying, cheating, and manipulation of others, but people would think about each other differently. Would you stand at the front of a subway platform if you knew that the other people there were thinking they should be on the lookout for opportunities to save lives by pushing you onto the tracks? How would you feel about crossing streets if you knew that others were always on the lookout for opportunities to save lives by throwing you into the path of an oncoming vehicle? Though he was mistaken about its not having any exceptions, Kant was right that the principle of not using others merely as means is an important moral principle. His own moral psychology prevented him from appreciating that it can be much more effectively implemented by a practice that simply "silences" certain thoughts about ways of treating people, rather than by a practice of consciously considering all the alternatives, including killing people to harvest their organs, and then applying the principle to rule out the alternatives that involve using other people.

In the consideration of cases in which the "using" of others involves killing them or letting them die, there are great benefits to a social practice that simply "silences" those alternatives in most cases. The main principle will permit exceptions, but only if the exceptions are so narrowly and carefully drawn that their effects on our way of thinking and responding do not spill over into ordinary cases. The trolley case exception can easily be extended to similar kinds of cases—for example, the pilot of a plane that has lost power who cannot avoid a crash, but can choose between a single family house and a busy school. But pushing large men off bridges to stop runaway trolleys is not in the same category. Setting aside the fact that such a practice would lower the life prospects of large men, the more serious consequence is that everyone, of whatever size, would come to be regarded as a potential trolley barrier or a potential runaway automobile barrier and that would greatly impact everyone's life prospects.

I conclude that the multiple-time-slice main principle, understood as a principle that evaluates social practices rather than norms or principles, is an improvement over Thomson's revised trade-off idea. But the main principle is not merely an improvement on one of Thomson's four exceptions to natural rights. It provides a way to subsume all four of them. Thomson's first two exceptions (word-giving and promising) correspond to what I have called the actual consent exception. It is easy to see how the main principle would justify these exceptions. Think of how greatly our life prospects would be diminished if we could not voluntarily waive our right not to be harmed for necessary medical treatment or if we could not enter into voluntary agreements. Of course, as I have previously explained, the main principle endorses exceptions to these exceptions also.

The main principle also subsumes Thomson's third exception, the exception for legitimate government action, especially through laws. The main principle plays this role, because it endorses some coercive laws even if

we assume a state of nature in which there is a prohibition on coercion. And finally, the main principle supersedes Thomson's revised trade-off idea, for the reasons just discussed. So the main principle subsumes all four of Thomson's exceptions to natural rights. It is the only principle we need to explain all of her exceptions to natural rights norms.

Security Rights as Rights to Legal Protection

Security rights are rights to legal protection against basic harms. Thus, they include both substantive rights and procedural rights necessary to protect the substantive rights. Some authors have limited security rights to rights against the government or other authoritative institutions (e.g., Pogge 2000). There is some basis for this in ordinary language, for we think of government torture or genocide as a human rights violation, but we do not usually think of a single murder or a single instance of spousal abuse as a human rights violation. I think that the explanation for this distinction is that a single instance of murder or spousal abuse is not a human rights violation if the government has in place an effective system for deterring and punishing murder and spousal abuse. If not, if the government allows some groups to murder with impunity or if the government does not even regard spousal abuse as a crime, then these abuses rise to the level of human rights violations.[18]

SEVEN

Epistemological Foundations for Human Rights

> There is the greatest difference between presuming an opinion to be true, because, with every opportunity for contesting it, it has not been refuted, and assuming its truth for the purpose of not permitting its refutation.
>
> —J. S. Mill

> Experience has proved that allowing a free flow of ideas can improve stability and alleviate social problems.
>
> —2/2/2006 open letter from former Communist Chinese officials and scholars, including a former secretary to Mao and a former party propaganda chief, criticizing government censorship

In the first volume (Talbott 2005), I emphasized the role of autonomy rights in solving an epistemic problem, the reliable feedback problem, which is the problem for governments of obtaining reliable information about how effectively their policies promote the well-being of citizens. In this chapter I generalize that discussion to consider the role of autonomy rights in solving a more fundamental epistemic problem—the problem of generating knowledge and rational belief, not only in individuals, but in the aggregate. If the rationale for autonomy rights is the equitable promotion of life prospects, their ground is in the epistemic limitations of individuals and in the possibility of social processes of rational belief and knowledge formation. In this chapter I lay the epistemological foundation for the explanation of the contours of the autonomy rights, especially freedom of expression and freedom of the press, in the next chapter.

The question that I ultimately want to answer is this: Why would a consequentialist principle like the main principle support *robust* rights to freedom of expression and to the other autonomy rights—that is, rights that a government could not justify infringing simply because the government thought that the infringement would better promote (appropriately distributed) well-being and rights that could not be overridden by a simple majority? I postpone the discussion of *inalienability* to chapter 10.

If, following Rawls, we think of the legislature as the arena in which decisions are made about how to best promote well-being, the challenge to the consequentialist is to explain why the legislature should not be able to

trade off restrictions on autonomy rights for gains in (appropriately distributed) well-being. Note that the nonconsequentialist has no problem answering this question, because the nonconsequentialist does not justify rights by their contribution to (appropriately distributed) well-being. But the consequentialist has some explaining to do.

In chapter 4, I reviewed Mill's first, unsatisfactory, answer to this question. Mill claimed that individuality (his term for autonomy) was one of the essentials of well-being. Logically, this seals the case. If autonomy is an essential of well-being, then the legislature would have to guarantee whatever was necessary for autonomy just to be able to promote well-being at all. We should be suspicious of any argument that makes the case for autonomy rights this simple. In chapter 4, I explained why Mill's argument fails. Now I turn to the early Rawls's attempt to fill the gap.

Rawls on the Priority of the Basic Liberties

What I am calling autonomy rights are very similar to Rawls's *basic liberties*. Rawls explained the priority of the basic liberties as the *lexical priority* of the liberty principle over the difference principle in his special theory of justice (1971, 302–303). The idea was that certain guarantees of equal basic liberties were so important that such liberties could be limited only in order to strengthen the overall system of equal basic liberty, not merely in order to promote well-being (represented by expectations of primary goods), not even the well-being of the least advantaged group (302).

In *Theory of Justice*, the early metaphysical Rawls thought that he could provide a consequentialist explanation of the lexical priority of the liberty principle. The idea was to show how his special theory of justice could be derived from a wholly consequentialist general theory of justice. The general theory contained a single social practice consequentialist principle, the maximin expectation formula (1971, 303).

The early Rawls did not think that the derivation of the special theory from the general theory of justice would go through in all circumstances. He limited the application of his special theory to societies that had reached a level of development that made it possible to effectively establish the basic liberties (1971, 152). Thus, the goal was to explain why the special theory, including the lexical priority of the liberty principle over the difference principle, would apply to all societies that had reached the necessary minimal level of development.

How did the early Rawls think that he could derive the lexical priority of the liberty principle from his general theory of justice? He introduced a fudge factor, which I refer to as the *value of liberty fudge factor*. The early Rawls simply argued that after society reaches a certain level of development, it cannot be rational to trade any basic liberties for other things that one desires, or at least that such trade-offs would be irrational in the original position

(1971, 151–152 and 542–548). Because the early Rawls based the argument on the irrationality of trading basic liberty for other kinds of goods, it was bound to fail. It is the fact that it *can* be rational to trade a basic liberty for other kinds of goods that makes such trades a collective action problem. In chapter 10, I explain why regarding the basic rights as inalienable is a solution to a collective action problem.

In any case, the later Rawls disavowed this argument for the lexical priority of the liberty principle over the difference principle and, in addition, gave up the idea of providing a consequentialist explanation of the priority of the basic liberties (1993, 371 n. 84). For the later Rawls, giving up on this project required him to give up the claim that rights to the basic liberties are a universal requirement of justice. Instead, the later Rawls replaced the metaphysical project of the early Rawls with the political project of trying to explain the importance of the basic liberties in democratic societies.

Early and late, Rawls has been clear on what the point of guaranteeing the priority of the basic liberties would be: the development and full and informed exercise of the two moral powers, rationality (the capacity to form, revise, and pursue a life plan) and reasonableness (the capacity to understand, apply, and be motivated by fair terms of social cooperation) over a complete life—in short, *full autonomy* (1993, 293, 302). The challenge is to provide a consequentialist explanation of the priority of the rights that are necessary for full autonomy—that is, *autonomy rights*. The gap in his account is to explain how a principle that evaluates practices in terms of equitably promoting well-being could endorse giving priority to the rights that are necessary for the development and full and informed exercise of the two moral powers.

Prospects for filling the gap look bleak. And yet it can be filled. Indeed, Mill himself showed us how to fill it. Unlike Mill's first answer, which gave the desired result directly, though implausibly, Mill's second answer requires some time to explain. The explanation begins with a revolution in epistemology.

Mill's Revolutionary Epistemology

In the first chapter of *On Liberty*, Mill announced that he was going to argue for what I am calling *autonomy rights*, on the basis of their contributions to utility (well-being). So, when in chapter 2, he takes up his defense of a right to freedom of thought and discussion, we expect him to justify it on the basis of its contribution to utility. However, in chapter 2 Mill makes practically no reference to utility. He discusses how freedom of thought and opinion contributes to the development of true belief, rational belief, and ultimately knowledge. In doing so, he presents a new and revolutionary epistemology. Mill's epistemology is a crucial part of his consequentialist justification of autonomy rights, so I need to say something about it.

Mill's revolutionary epistemology was not well understood by his contemporaries, because he was proposing a complete reorientation of epistemology from its preoccupation with the reasons accessible to an individual from the inside, which I refer to as *individualistic epistemology*, to an account of justification and knowledge as the product of a certain kind of social process. This idea of justification and knowledge as a social process would ultimately become extremely influential in twentieth-century philosophy, but by then, Mill's emphasis on identifying processes that lead to truth (or, at least, better approximations of it) would seem almost quaint. Ironically, for all their differences, both the pragmatists in British-American analytic tradition and the postmodernists and deconstructionists in the Continental tradition converged on the conclusion that because our beliefs have whatever kind of justification they may have in the context of a contingent social process, there is no role to be played in epistemology by context-independent or context-transcendent notions such as truth, because we would have no access to it. Call this the *immanence of thought* hypothesis: All rational thought involves immanent not transcendent concepts or ideas.

It was not until late in the twentieth century that the pendulum began to swing back toward Mill. It is useful for me to briefly recapitulate the history, beginning with Mill.[1]

Mill was one of the first to articulate the new epistemology, because he was one of the first to completely reject the main foundation of the old epistemology, *a priori* justification. Mill's complete rejection of the *a priori* may have been too extreme, but it enabled him to ask what would be the central question in philosophy in the twentieth century: If philosophy itself cannot be done *a priori* (and the record of mistakes and disagreements in philosophy made it clear that it could not), how could it be a rational enterprise and how could it aspire to truth and knowledge? Because Mill completely rejected the *a priori*, he asked this question about every kind of knowledge—logic, philosophy, mathematics, science, history, political theory, and morality. In every area of inquiry, he recognized the same kind of process for rationally testing beliefs and making progress. Having seen the process work in so many different areas of inquiry, in chapter 2 of *On Liberty*, he described it. It was a process that depended on the free expression of opinion, even if the opinion was generally regarded to be mistaken, for several reasons:

(1) *Fallibility*. Any of our beliefs might be mistaken. Because of his rejection of *a priori* justification, Mill was prepared to admit we might be mistaken about anything.[2]

(2) *Portions of the truth*. Mill argued that when we find a disagreement, it is rare that one side is completely correct and the other side is completely mistaken. Each side typically has some portion of the truth. Our ability to make progress toward the truth depends on the possibility of all sides being able to express themselves, so that at least impartial observers can increase the portion of truth in their opinions and get closer to the truth.

(3) *Rationality*. Even if their opinions were false, those whose opinions conflicted with a dominant position would perform a valuable function. Because people give reasons for their opinions, those who disagreed with the dominant view would present reasons that the defenders of the dominant view would be challenged to respond to. In this way, the dominant view would continue to be rationally held and not become mere dogma or prejudice. Before Mill (and even after), philosophers thought that their goal was to end disagreement by articulating a view that no one could rationally disagree with. With his rejection of *a priori* justification, Mill gave up this view of philosophy. In Mill's epistemology, if everyone agreed with his philosophy, that would end the process of rational inquiry and thus end any hope for more progress toward truth.

Thus, Mill argued, rationality and progress toward the truth in any area depend on there being a process of *the free give-and-take of opinion*. It is important to appreciate just how radical this proposal was. In what seems now like an act of prescience, Mill used Newtonian physics to illustrate his epistemology. At the time, Newtonian physics was regarded as rationally unquestionable. Kant had argued that it was synthetic *a priori*. In the nineteenth century, it was regarded as established beyond doubt. Mill demurred:

> If even the Newtonian philosophy were not permitted to be questioned, mankind could not feel as complete assurance of its truth as they now do. The beliefs which we have most warrant for, have no safeguard to rest on, but a standing invitation to the whole world to prove them unfounded. If the challenge is not accepted, or is accepted and the attempt fails, we are far enough from certainty still; but we have done the best that the existing state of human reason admits of; we have neglected nothing that could give the truth a chance of reaching us: if the lists are kept open, we may hope that if there be a better truth, it will be found when the human mind is capable of receiving it; and in the meantime we may rely on having attained such approach to truth, as is possible in our own day. This is the amount of certainty attainable by a fallible being, and this the sole way of attaining it. ([1859], 28)

Mill's contemporaries could not appreciate such a radical proposal. What he said about Newtonian physics, which was radical in his day, is commonplace today. The subsequent replacement of Newtonian physics with relativity theory and quantum mechanics have made Mill look prescient.

The significance of the process of the free give-and-take of opinion for justification and knowledge is better appreciated today than it was in Mill's day. However, there are still many misconceptions about it, including one of Mill himself. In order to address these misconceptions, I trace out some of the main developments in the historical development of a social process theory of justification and, thus, of knowledge.

Peirce's Ideal Social Process Theory of Truth

The next person to make such a radical proposal was C. S. Peirce, one of the founders of pragmatism. Like Mill, Peirce was a fallibilist who rejected *a priori* justification. However, Peirce did not accept the move that came to characterize pragmatism, to make truth itself immanent to the process of inquiry (e.g., James [1897]). For James, a true belief was just one that worked. Peirce objected to metaphysicians who made truth too transcendent for human beings to access it, and he objected to pragmatists like James who made it so accessible as to not require serious inquiry. So Peirce tried to find a compromise theory of truth as immanent-transcendent. His compromise was to define truth as the ideal limit of rational inquiry (which, for Peirce, was scientific inquiry), a limit that could never actually be reached, but only approximated (Peirce 1992/1998). That makes his theory of truth an *ideal social process* theory.

Peirce's *ideal social process* definition of truth was intended to apply only to scientific inquiry, not normative moral or political inquiry. I refer to this as his *rejection of normative transcendence*. In this, he was part of a broad consensus that spanned borders and traditions, for it was dominant in both Anglo-American and Continental philosophy for most of the twentieth century. On the Continent, Marxism, existentialism, hermeneutics, poststructuralism, postmodernism, deconstructionism, and critical theory coincided on this issue, if on nothing else, with logical positivism, pragmatism, and naturalism in Anglo-American philosophy.[3] Their rejection of normative transcendence made them all relativists in one way or another about normative moral and political inquiry.

The Early Metaphysical Rawls

Thus, the year 1971 was an important year in Anglo-American political philosophy, because the publication of *Theory of Justice* heralded the possibility of inquiry into universal truths of justice. The early Rawls of *Theory of Justice* showed how to replace traditional foundationalist epistemology with a fallibilist epistemology based on reflective equilibrium. In a reflective equilibrium model, we try to bring our moral principles into reflective equilibrium with our considered moral judgments about particular cases. Thus, Rawls provided a model of moral inquiry that did not depend on *a priori* insight into self-evident principles, and indeed made moral theory a kind of explanatory theory.

It is true that the early Rawls's epistemological model was individualistic rather than social, but it was easily extended to a social model, and he himself so extended it with the introduction of full reflective equilibrium (1995, 141 n. 16). Thus, the early metaphysical Rawls provided an epistemology that would explain how we could be a part of a social process that, through

the use equilibrium reasoning, could discover true principles of justice, or at least to make progress toward discovering them.

The early Rawls not only provided the epistemology for a theory of justice, he actually claimed to be able to tell us what the principles of justice would be for an ideally just society.[4] On my reconstruction, what made *Theory of Justice* a metaphysical theory is that early Rawls was presenting arguments that he expected to be persuasive to anyone in any cultural tradition: first, that fairness requires them to be willing to cooperate on principles of justice that would be accepted in the original position, behind the veil of ignorance; and, second, that in the original position everyone would agree to his two principles of justice (including the lexical priority of the first over the second). In the ideally just society envisioned by Rawls, there would be lots of reasonable disagreement, but there would be universal agreement on the two principles of justice, because everyone in that society would see that there could be no reasonable basis for rejecting them. This idea of finding principles on which there could be no reasonable disagreement became the defining mark of Rawls's philosophy and has been probably the most influential idea in the philosophy of human rights. However, as I explain shortly, it is a serious mistake.

Shortly after publishing *Theory of Justice*, Rawls came to realize (perhaps as a consequence of the deluge of critical responses to *Theory*) that it was a mistake for him to insist that there could be no reasonable disagreement with his two principles or indeed with the liberal conception of justice. At this point, he faced a momentous choice between two projects, metaphysical liberalism and political liberalism:

(1) *Metaphysical liberalism*. Continue to advocate a liberal conception of justice as a universal theory of justice. To take this route, he would have had to argue that those who advocated nonliberal forms of government were committed to principles that they could not justify impartially in the original position, behind the veil of ignorance, and that that was a substantive reason for rejecting their principles.

(2) *Political liberalism*. Relativize his theory to those who accept liberal principles of justice. If he had relativized the theory to those who explicitly accepted his two principles, it would have been relativized to a very small group. Instead, Rawls relativized the theory to those whom he saw as implicitly accepting liberal principles, if not exactly the two favored by Rawls himself.[5] In his terms, the new goal would be to articulate "certain fundamental ideas seen as implicit in the public political culture of a democratic society" (1993, 13).

As is well known, Rawls (1985) chose the second course, political liberalism, and disavowed metaphysical liberalism. This move to political liberalism gave the terms *reasonable* and *unreasonable* an entirely new meaning. Whereas in metaphysical liberalism, *unreasonable* would have meant something like *not willing to acknowledge reasons that apply impartially to everyone*, in political liberalism, *unreasonable* as applied to comprehensive views,

simply means *not part of an overlapping consensus on a liberal conception of justice* or equivalently in Rawls's account, *not willing to be bound by the results of the original position* (1993, 62; 1999, 87). There is no longer any implication that someone who was not willing to accept a liberal conception of justice is making a mistake about something (i.e., justice) or failing to recognize the force of reasons that should be recognized by everyone. To emphasize this point, in *Political Liberalism* Rawls explicitly rejected any claims to *truth*. In talking about people's comprehensive moral, political, philosophical, and religious doctrines, Rawls replaced "true" with "reasonable" (128). Nothing in *Political Liberalism* is intended to imply the falsity of *any* normative moral and political view. *Political Liberalism* intentionally leaves it open whether an *unreasonable* doctrine such as Nazism is true. And though *Political Liberalism* classifies Nazism as unreasonable, all that amounts to saying is that Nazism does not accept a liberal conception of justice.

Thus, though Rawls continues to use the term *unreasonable* in *Political Liberalism*, the term has been relativized to the public culture of democratic society. It would have been more perspicuous if Rawls had introduced a new term *unreasonable*$_L$ and used that in place of *unreasonable* throughout. To see this, suppose that a new Hitler wrote a new defense of Nazism, *Political Nazism*. If the new Hitler claimed to be articulating the fundamental ideas seen as implicit in the public political culture of a Nazi society, he could specify the basic principles of Nazi justice and define *unreasonable*$_N$ as *not willing to accept the basic principles of Nazi justice*. It can come as something of a shock to realize that nothing in *Political Liberalism* provides the slightest reason for thinking that being unreasonable$_N$ is any more or less unreasonable, in the unqualified sense, than being unreasonable$_L$.[6] This is not true of *Theory of Justice*. In *Theory*, the early metaphysical Rawls would have argued that Nazism was unjust, not merely unreasonable$_L$, because its principles could not be agreed to in the original position. Political Rawls defends a relativist position on reasonableness and on justice.

Habermas's Ideal Social Process Theory

When Rawls took his political turn, it became clear that normative antirelativism would need another champion. Fortunately, a champion appeared, Jürgen Habermas. Habermas insisted that rational inquiry was a cooperative social process, which he referred to as *communicative action* (i.e., "communication oriented to reaching understanding" [1993, 59]) or *rational discourse*. Habermas pointed out that although discourse takes place in a contingent context (a lifeworld), when we engage in it we come to recognize that it commits us to a concept of validity that transcends our particular lifeworld and, indeed, all particular lifeworlds. Habermas insisted that all discourse (understood as communicative action), moral as well as scientific, carries this commitment to transcendent validity. Whenever we engage in

discourse, we commit ourselves to the transcendent validity of the results of an ideal process of the free give-and-take of opinion (which he referred to as an "ideal speech situation" [1990, 88]), governed by various norms, both moral and nonmoral. Habermas identified the four most important norms: (a) publicity and inclusiveness—the discourse is open to all; (b) equal rights to engage in communication; (c) exclusion of deception and illusion; and (d) absence of coercion (2003, 106–107). Communicative action is a cooperative activity in which the goal is understanding and the only force that may be used is the force of the better argument (1990, 87–89).

Initially, Habermas, like Peirce, was worried about the metaphysical implications of transcendent *truth*, so he adopted a Peircean ideal process definition of both the validity of the purely descriptive and the validity of the normative, as the end product of this ideal process of discourse. So that the validity of the purely descriptive would not be confused with the validity of the normative, he reserved "true" for the validity of the purely descriptive. However, initially, his accounts of both purely descriptive and normative validity were epistemic—that is, accounts in terms of the output of the ideal process of discourse.

Over time, Habermas came to the conclusion that an epistemic conception of truth, (i.e., a theory of truth as the result of the ideal process of discourse) was not adequate as a theory of purely descriptive validity. So he replaced the epistemic conception of truth with an objective one (2003, 91–92). His reason was simple: On purely descriptive questions, a proposition is agreed upon in rational discourse because it is true; it is not true because it could be rationally agreed upon (2003, 101). Call this the *Euthyphro test* for concepts of validity. I employ a version of this test in chapter 10.

Although the Euthyphro test led Habermas to an objective conception of purely descriptive validity, he continues to hold an epistemic theory of normative validity in terms of the ideal process of discourse. "The meaning of 'moral rightness,' unlike that of 'truth,' is exhausted by rational acceptability" (2003, 109). This is because valid norms are aimed at creating something—"legitimately ordered interpersonal relations in the social world" (2003, 54).

Habermas reserves the term *moral* for norms, including but not limited to norms of justice, that presuppose this sort of lifeworld transcendent validity. Habermas thinks it is a presupposition of declaring something to be a moral norm or a norm of justice that it "must be able to command the rationally motivated recognition of *all* subjects capable of speech and action, beyond the historical and cultural confines of any particular social world" (2003, 104; emphasis in original) He contrasts moral norms with *ethical* norms, which do not transcend lifeworlds, because they concern what is necessary to live a good life within a given lifeworld (1993, 5).

If the same ideal process of rational discourse is presupposed by both purely descriptive and moral validity, then moral discourse must be distinguished from scientific discourse by the subject matter of the discourse.

Habermas believes that moral norms are the norms that emerge from the ideal process of discourse when it is directed toward agreement on norms that satisfy the following condition: "(U) All affected can accept the consequences and the side effects its *general* observance can be anticipated to have for the satisfaction of *everyone's* interests" (1990, 65). In the process of discourse, the goal is not for each to bargain for his greatest advantage, but to try to determine what norms would be "equally good" for all affected (2003, 33–34).

Habermas's commitment to lifeworld transcendent validity has made him an opponent of almost every kind of relativism, moral and nonmoral, whether in the Continental or Anglo-American philosophical tradition. He has been one of the leading critics of Rawls's move to political liberalism, because of Rawls's failure to recognize the lifeworld-transcending presuppositions of normative validity claims such as claims about what is *reasonable*, (1995, 124). In the terms discussed above, on Habermas's view, there is no reasonableness$_L$ or reasonableness$_N$, there is just reasonableness. Habermas's criticism of all these forms of relativism has two parts: first, that they all involve what Apel (1988) refers to as *performative contradictions*, because in arguing for relativism of any kind, one presupposes the lifeworld transcendent norms of rational discourse; second, going beyond Apel, Habermas argues that in engaging in rational discourse, one presupposes the lifeworld transcendent validity of the results of the ideal process of discourse defined by those very norms (1990, 80–81).

Given the influential role that Habermas has played in opposing relativism, it is somewhat surprising that I will suggest that his own position is relativist in a problematic way. Given his criticism of Rawls for moving to political rather than metaphysical liberalism, it is surprising that there is a parallel criticism of his own view. As I explain shortly, when forced to choose whether to interpret the commitments of his own theory as metaphysical or factual, he chooses factual. Thus, for both Rawls and Habermas, fear of metaphysics leads to relativism.

Habermas's Presuppositions of Rational Discourse: Factual Not Metaphysical

Suppose that Habermas is correct that in participating in communicative action we must presuppose the lifeworld transcendent validity of the results of the ideal process of discourse. The obvious question is whether that presupposition itself is rational. If it is rational, it is hard to believe that what *makes* it rational is that it would be agreed upon by the participants in an ideal process of discourse.

Habermas is aware of this question. He believes that the only way to answer it positively is to accept what he calls Kant's "illusion of pure reason" (2003, 83). If we had pure reason, then pure reason could tell us that we *must* make accept these presuppositions to be rational. Without pure reason,

Habermas thinks that when we recognize that we *must* accept these presuppositions, the *must* "does not have the transcendental sense of universal, necessary, and noumenal" (2003, 86) but rather the inevitability of what is *inescapable* for us, where the inescapability is factual. There is no alternative for us to our kind of communicative action (1990, 89, 94, 102, 116; 2003, 85–86).

As I explain shortly, I think this claim is false, but let's suppose that it were true. We could still ask if it is rational to accept such presuppositions. It is easier to accept the fact of inescapability if what is inescapable makes us rational rather than irrational. However, even to suppose those presuppositions were rational would seem to commit us to there being some objective normative truths—in this case, truths about what norms are *really* rational for us, not merely inescapable. Thus, it is natural for Habermas to suppose that to discern such truths we would have to have a faculty of pure reason. But this is a mistake; there might be other ways of discovering which norms were really rational for us than by an individualistic faculty of direct rational insight. Perhaps the social process of rational inquiry would be a way of figuring out what they were.

In any case, by making lifeworld transcendent validity a product of presuppositions that are merely inescapable for us, Habermas makes his account relativist. It applies to beings for whom, like us, certain presuppositions are inescapable. Call those the *presuppositions of human discourse* (*presuppositions of discourse*$_H$). Just as there is nothing in Rawls's political liberalism that even bears on whether Nazism is the true moral theory (i.e., on whether reasonableness$_N$ rather than reasonableness$_L$ is the true theory of reasonableness), there is nothing in Habermas's theory that even bears on whether it is possible that there be beings for whom the presuppositions of discourse$_H$ are not inescapable, and, if so, whether the results of ideal discourse$_H$ have any kind of validity, purely descriptive or normative, for them.

Habermas might accept this implication, for it would seem that in order to evaluate ideal discourse$_H$, we would have to be able to step outside the presuppositions of discourse$_H$, which, by hypothesis, are inescapable for us. However, I think there is another alternative. All that is necessary for the results of ideal discourse$_H$ to have universal validity is for there to *be* some universal standards of rationality according to which the presuppositions of discourse$_H$ are rational for human beings. This is a metaphysical condition. Because universal standards of rationality would be true in all possible worlds, they would be necessarily true.

It is important to separate this metaphysical question from the epistemological question of how we could come to rationally believe in the existence of necessarily true standards of rationality. There is a tendency to think that the only way to acquire a rational belief in the necessity of some proposition is to have a faculty of pure reason to directly access necessary truths. This is a mistake. When we think of moral reasoning as bottom-up rather than top-down, we open up the possibility that the proposition that certain truths or

moral norms are necessary could be justified by *its* explanatory role. The fact that certain truths are necessary would be required to explain not only all actual cases but also all hypothetical cases. What I am suggesting is that Habermas should not have been so quick to prejudge the status of the presuppositions of human rational discourse. He should have been willing to wait and see what verdict the process of rational discourse would give on the metaphysical status of its presuppositions.

Another consequence of the possibility that I am outlining is that the model of descriptive validity as objective truth might also extend to the moral realm. Habermas is correct to insist on important differences between purely descriptive validity and normative validity. Norms are not objects that we can bump into or physical laws that we can test with experiments. But that does not mean they are not objective. In denying that there are objective norms, Habermas is siding with the constructivists and proceduralists against the substantive normative realists. Constructivists and proceduralists hold that the valid or true norms are those that are (or would be) selected by some process P. There is no objective normative truth or objective normative validity; there are only the norms that are (or would be) selected by process P.

The problem with constructivist and proceduralist accounts is to explain the status of the following normative statement:

> (P-Norm) The true or valid norms are those that are (or would be) selected by process P.

It does not seem that the truth or validity of the P-Norm could be explained by its being selected by process P (even if it were). We have seen how Habermas responds to this problem. He simply claims that his version of the P-Norm is an inescapable presupposition of human discourse. How ironic it would be if the process of human discourse came to a different answer to that question.

Inescapability Is Not a Fact

Thus far, I have been assuming for the sake of argument that the norms of human discourse really are inescapable for us. Now I argue that this is not true. This will have an important consequence for what I call the *priority* of the ideal process of discourse.

Let me show you how you could cease to identify validity with the results of ideal human discourse. Suppose that one day you come upon a burning bush and a voice announces to you that it is the voice of God. At first you are skeptical, but over the course of many conversations, you become convinced. At first, the voice just correctly predicts the outcomes of horse races and other sporting events. You soon begin wagering on the voice's predictions and quickly become quite wealthy. Then the voice tells you how to perform

an experiment that will disprove general relativity. You use your newfound wealth to fund the experiment, and when you publish the results, you are awarded the Nobel Prize in physics. Imagine as much additional confirming evidence as you would need to convince you that this voice really is omniscient. Then, one day, the voice tells you the following: The true laws of the universe are deterministic, but human inquirers lack the conceptual resources necessary to understand them. The best that human beings will ever be able to do is to formulate probabilistic laws that are approximations of the true deterministic laws of the universe. In addition, the voice informs you that our universe did indeed begin with a Big Bang and that no ideal process of human discourse would ever be able to determine what happened before the Big Bang. Then the voice tells you what happened.

At some point in this story, I would expect you to have given up your presupposition that the results of the ideal process of human discourse are true or, at least, that truth is identified with the results of that process.[7] You would have found a superior process for determining what is true. That would detach purely descriptive truth from the ideal process of discourse.

What about normative validity? Let us suppose that the voice gave you surprising new arguments to resolve many outstanding questions in moral theory. When you publish the articles, they are received with general acclaim. The voice also gives you advice on settling the Israeli-Palestinian conflict and other conflicts around the world. The advice is successful. As a result, you win the Nobel Peace Prize. Then the voice tells you this: There are true, meta-level consequentialist principles that explain what the ideal norms of human morality and human justice should be. Unfortunately, the meta-level consequentialist principles are so complex that no human being could ever understand them. Even in an ideal process of discourse, the best that human beings could do would be to agree on norms that approximated what the norms of justice should be. As a result, human societies could never be perfectly just, even if they satisfied the norms that would be agreed on in an ideal process of discourse.

I believe that in this case, it would be reasonable for you to give up the presupposition that normative validity is determined by the ideal process of discourse. How would Habermas respond? I think he would have to be committed to saying that there is no right answer to moral questions in advance of carrying out the process of discourse. What is the status of that claim? Habermas could not think that it was justified *a priori*. So it must itself be open to discussion in the process of the free give-and-take of opinion.

The example of the voice from the burning bush is meant to break the connection between the ideal process of discourse and validity. Once we do so, we are in a position to see that Habermas has misunderstood the significance of the ideal process of discourse. It is not the ideal process of discourse that determines the validity of our ordinary process of discourse. Quite the contrary, it is through our ordinary process of discourse that we are able to identify ways of improving that very process, and thus to fallibly articulate

an ideal of rational discourse. Our conception of an ideal process of inquiry is just our best attempt to characterize the kind of process that would be best for getting at the truth. This ideal is itself constantly evolving as we engage in our ordinary process of discourse and discover impediments to getting at the truth and think about how to avoid them.

In Habermas's account, the ideal process has a kind of logical priority over our actual process. Our actual process of inquiry achieves validity only to the extent that it approximates the ideal process. I believe that we should reverse the logical priority. Our belief that the ideal process is itself ideal depends on the validity of our actual process, because our conception of the ideal process is itself a result of the actual process of discourse. If our actual process of discourse is unreliable, it is probably unreliable about the properties of an ideal process of inquiry also.

Once we reverse the priority that Habermas assigns to the ideal process, we can solve another puzzle for Habermas's view. How do we determine the norms of ideal inquiry? If Habermas is correct, to identify them, we would have to articulate only our own inescapable presuppositions. But this is not how we determine the norms of ideal inquiry. We figure them out by a process of discovery. We discover the kinds of factors that impair our own process of discourse from achieving valid results, and then we try to articulate norms that rule out those factors.

For example, Habermas's own early lists of the norms of ideal discourse included a norm against lying, but not against self-deception (a least if the self-deceiver *believes* what she is self-deceived about),[8] because his norms permitted participants in ideal discourse to say whatever they believed. In later writings, he has interpreted sincerity as a requirement "to be honest with oneself" and "to critique one's self-delusions" (2003, 269). But there is no way, from the inside, of determining whether one's beliefs are due to self-deception or bad faith. So there is no way of being able to tell from the inside if one is complying with the norms of ideal discourse.

Moreover, the example of self-deception shows that the norms themselves have to be discovered by the process of the free give-and-take of opinion. I suspect that people engaged in discourse for centuries before they even discovered self-deception and bad faith. Almost any way of discovering them will involve recognizing that they bias thought and, ultimately, the process of discourse in a way that is an impediment to achieving valid results. Only then would people be in a position to formulate a norm prohibiting self-deception and bad faith. Before they were discovered, no one could have presupposed a norm prohibiting them.[9]

On reflection, it is hard to see how Habermas could have thought that we could identify the ideal process of discourse merely from our own presuppositions. Note that Habermas's own ideal of discourse is based on argumentation. It is not surprising that a philosopher would identify the ideal form of discourse with argumentation. However, it is easy to imagine a time in the past when people's ideal of discourse would not have been a form of argumentation.

And it is also not difficult to imagine that people's ideal in the future won't be a form of argumentation either. It could be, for example, that human beings are on a trajectory of improving their discourse practice and that argumentation is only a middle stage of the trajectory. This seems to me to be especially true of moral discourse. When we imagine a group of people trying to work out the norms that would be equally good for all, it might well be that a process of role-playing and sharing life stories and aspirations might be a better process than any process of argumentation for working out norms acceptable to all.[10] Of course, I could be mistaken about this. The important point is that our conception of the ideal process of discourse does not have to be defined by whatever norms we now presuppose as ideal. We can allow for our opinions about the ideal norms to evolve over time as we understand better the kinds of impediments to success in rational inquiry.

If this is right, then our conception of ideal discourse is just one more of our beliefs that is evolving as a result of our ordinary process of the free give-and-take of opinion. Imperfect and qualified though it is, it is that ordinary process that enables us to form and improve on all our views, including our conception of ideal discourse. In Neurath's famous image, we must rebuild our raft while using it to keep us afloat. If we have a conception of an ideal raft, it too is subject to revision in light of what we discover about rafts as we rebuild and improve our less-than-ideal one.

And thus it is that at the beginning of the twenty-first century, we are now in a position to fully appreciate Mill's revolutionary claim about the importance of the ordinary—not ideal—process of the free give-and-take of opinion in epistemology: Our ability to apprehend at least partial or approximate truths and to have more or less rational beliefs in science or morals or any other area is typically due to our participation in a real, not hypothetical, process of relatively free give-and-take of opinion. However, there is one ingredient to be added that was missing from Mill's own epistemology, as it was missing from Peirce's, Rawls's, and Habermas's. Because we don't have direct rational insight into necessary truths, the free give-and-take of opinion is also our only way of attaining a rational belief on whether some normative truths are strictly universal—that is, true of all rational beings in all possible worlds—and, if so, which ones. The process of the free give-and-take of opinion could lead to general recognition of some necessary truths of this kind. In any case, there is no need to *assume* that some form of relativism about the purely descriptive or the normative is true. The process of free give-and-take of opinion is still addressing that issue. The jury has not issued a verdict.

We are now in a position to understand what I call the *Millian epistemological argument* for a right to freedom of expression and the other autonomy rights—*Millian* because it is a development of Mill's argument. I take up this argument in the next chapter.

EIGHT

The Millian Epistemological Argument for Autonomy Rights

The process of free give-and-take of opinion is the foundation of all human rights, especially the autonomy rights. The autonomy rights are the rights that are necessary for the development and exercise of *autonomy*, understood in my consequentialist sense as the combination of *good judgment* (to be a reliable judge of one's own good) and *self-determination* (the capacity to have one's judgments guide one's actions). Traditionally, philosophers have thought of autonomy as an individual achievement. In this chapter I explain why it is a social achievement and focus on the epistemological preconditions for achieving it. Rather than discussing all of the autonomy rights, I focus on the two most important for attaining autonomy, rights to freedom of expression and to freedom of the press. Liberty rights against paternalism are discussed in chapters 12 and 13.

The Millian epistemological argument for freedom of expression and freedom of the press is that they are necessary to maintain and improve the process of the free give-and-take of opinion. The argument itself is a contribution to that process. The argument concerns how to maintain and improve the process—not from the point of view of an ideal process of discourse, which we could never access and the norms of which we can expect to discover only by engaging in the free give-and-take of opinion—but from within the very process. Finally, something that I return to shortly, the argument is offered with the understanding that it is almost surely mistaken in some respects and with the expectation that it will be improved by the process of the free give-and-take of opinion.

The Millian epistemological argument is not a complete consequentialist argument for freedom of expression, for even if the process of free give-and-take of opinion is necessary for rational belief and for progress toward truth, we still need to know what is so good about rational belief and progress toward the truth. I take up that issue later in the chapter. Also, the Millian epistemological argument is not an argument for freedom of all types of expression, only the expression of things that can be true or false. I continue to use *opinion* to refer to beliefs with propositional content—that is, beliefs that can be true or false. I show how Mill extended the argument to nonpropositional expression later in the chapter. One way to better understand the Millian argument is to consider potential misunderstandings of it. That is how I will proceed.

The Millian Epistemological Argument Does Not Cover All Expression

The Millian epistemological argument applies only to expression that has propositional content—that is, can be true or false—because only expression with propositional content could be part of a process of free give-and-take of opinion aimed at truth. However, because Mill believes that expression in any area of inquiry, normative (e.g., moral and political theory) as well as purely descriptive (e.g., science and history), has propositional content, the epistemological argument still supports a very broad spectrum of expression. Nonetheless, it does not cover art that has no propositional message (e.g., most abstract art and dance); nor does it cover fiction (e.g., literature and most drama), because fiction does not even attempt to say something true. This is not to say that Mill does not believe that these activities shouldn't also be covered by a right to freedom of expression, only that the argument will have to be an extension of the epistemological argument. The epistemological argument covers only the free expression of opinion.

It should also be mentioned that the argument does not apply to absolutely all opinions. Mill would not insist that absolutely every belief depends on the free give-and-take of opinion to be rational. Let us say of beliefs whose rationality does not depend on their being subject to the free give-and-take of opinion that they are *not dependent on the process*. Each of us has some personal beliefs, including the belief that we exist, that do not depend on that process to be rational, and there may be some apparent tautologies (e.g., that 2 + 2 = 4) that do not depend on it either. Because most personal beliefs make no significant contribution to the free give-and-take of opinion, it would be expected that Mill would allow for a privacy right that would support limitations on people's right to reveal personal information about others. I take up privacy rights in chapter 13.

As for apparent tautologies like 2 + 2 = 4, no one would ever consider suppressing them, so it does not matter that they are not dependent on the process. However, as Feinberg (1980) points out, there is one important category of beliefs that are dependent on the process: beliefs concerning where the line is to be drawn between beliefs not dependent on the process (e.g., 2 + 2 = 4) and beliefs that are dependent on it. This fact alone is enough to create a presumption against any restrictions on expression of personal beliefs or apparent tautologies.

Also, the Millian argument covers only the expression of opinions—things one actually believes. It does not cover deception ([1859], 19). It would be hard to give an epistemological justification of a process that encouraged lying. Thus, Mill's argument provides no protection for someone who falsely shouts "Fire!" in a crowded theater. When libel involves an intentional falsehood or when there is reckless disregard for the truth, it would definitely not be covered by a Millian right to freedom of expression. Whether an exception for libel would extend to negligent falsehood is not so clear to me. I think it

is clear that a Millian exception would not extend beyond negligent libel to cover nonnegligent libel for the simple reason that, evaluated as a practice, it would have a chilling effect on the free give-and-take of opinion.[1]

Nor does the Millian epistemological argument apply to most kinds of advertising. Much advertising has no propositional content. When it does, the propositional content is usually being delivered by people who have been paid to read a script, not to state their own opinions. There is another reason why advertising is not covered by the argument, which I postpone until after I discuss the question of a right to freedom of nonpropositional expression.

There is one extension of the argument that Mill himself makes in chapter 2 of *On Liberty*. After completing the main argument, he considers whether freedom of expression should extend to "invective, sarcasm, personality, and the like" ([1859], 62). Consider, for example, the political demonstrator who calls the police "pigs." The demonstrator has no intention of saying something true, so initially it seems that the epistemological argument does not cover this kind of expression. However, Mill argues that those who oppose the *status quo* often use this kind of strong language to express views with propositional content. To permit the suppression of the strong language would have the indirect effect of suppressing beliefs with propositional content that challenge the status quo. The importance of permitting the expression of those propositional beliefs makes it necessary to also tolerate the strong language used to express them.

It is an interesting question whether Mill would have made an exception for hate speech. I think a narrow exception can easily be justified, for derogatory speech directed at a particular person. However, it is clear that the exception would not extend to publishing derogatory articles in a newspaper or periodical or to publishing other opinions—for example, Holocaust denial, which is now illegal in several European countries. Besides interfering with the free give-and-take of opinion, such laws inevitably backfire by creating sympathy for those whose views are banned. When a government responds to dissent with force rather than with reasons, the unavoidable message is that the government does not think that reasons alone are strong enough to sustain its position.

The Millian argument is compatible with some restrictions on the expression of opinions with propositional content. The free give-and-take of opinion does not require that I be allowed to call you and give you my philosophical opinions any time of the day or night. All commentators on freedom of expression allow for reasonable non-content-based restrictions on expression. But Mill was prepared to allow content-based restrictions to avoid serious harm. It is easy to see that the main principle would justify exceptions to freedom of expression in cases in which there is imminent danger of serious harm, even if they were content-based. Mill gives the example of a speaker who, addressing an angry mob outside the home of a corn dealer, claims that corn dealers are starvers of the poor ([1859], 64). Mill had no objection to limiting

the expression of that opinion to avoid violence. What was important for Mill was that there not be any suppression that would exclude the opinion entirely from the free give-and-take of opinion. He insisted that the opinion that corn dealers are starvers of the poor should be protected when printed in a newspaper (64). Thus, though the main principle would justify some narrowly drawn content-based restrictions on freedom of propositional expression, the Millian argument creates a strong presumption against such restrictions and in favor of unrestricted expression, a presumption that is increased by the danger of a government's abusing the power to limit expression.

The example of the mob outside the corn dealer's house reminds us of a number of different categories of exceptions to freedom of expression that might well be classified as involving "clear and present danger of imminent harm," if that test had not been so reinterpreted by the U.S. Supreme Court in *Dennis v. U.S.* (1951)[2] as to apply to speculative and remote dangers. The current Supreme Court test is that to justify limits on expression, the harm must be imminent and probable.[3] This test can allow narrow exceptions for true military secrets, incitement, causing panic, and fighting words, none of which can be thought of as making significant contributions to the free give-and-take of opinion (cf. Feinberg 1980).

The Millian Epistemological Argument Is Not a Relativist Argument

The argument for the free give-and-take of opinion does not assume or imply that all opinions are equally valid. What makes it such a powerful argument is that it can be given by someone like Mill, who consistently and visibly insisted that all opinions are *not* equally valid. The argument is that even acknowledging the very great differences in the validity of individual opinions, the social process of the free give-and-take of opinion is the best way of improving them. As Mill points out, the process does not necessarily improve the opinions of the partisans in a particular controversy, but it can greatly improve the opinions of impartial observers ([1859], 60).

Also, *pace* Rawls, the argument is not limited to those who are committed to the fundamental ideas implicit in "the public political culture of a democratic society" (1993, 13), and, *pace* Habermas, it is not limited to those for whom the presuppositions of ideal discourse are inescapable. For Mill, it is a fact that human beings have no voice from a burning bush to give us direct insight into truth, so it is just a fact that our beliefs are fallible and the social process of the free give-and-take of opinion is our only way of having rational beliefs and of reliably making progress toward the truth.

It is also important to distinguish the Millian position from Fish's (1994), which it superficially resembles. Whereas Mill defends freedom of expression by its contribution to the free give-and-take of opinion, Fish regards any doctrine of freedom of expression, except the "pure" doctrine of no restrictions

on expression at all, as a political position that each person, as a voter, has a right to try to legislate in the political (not epistemological) process of democratic politics. As Fish would say: It's all politics. Mill would not accept that. He would hold that the epistemological process of the free give-and-take of opinion has priority over the democratic process, because the democratic process needs a source of reliable beliefs on which to base legislation. This turns out to be important, because it provides the basis for a logic of robust rights—that is, rights that should be protected against majority opinion. Majority opinion should not be allowed to override the rights that are necessary to make it rational. Fish could hardly make sense of such a position, because, for him, the label *rational* would itself be just another political slogan—at best, a strategy for trying to win over a majority to one's own view.

The Millian Argument Is Not a Skeptical Argument that the Government Is Unreliable in Determining Truth and Falsity

It is sometimes claimed that Mill thought that, at least on issues on which there is disagreement among citizens on the truth, the government could never be justified in making a determination in favor of the opinions of some and against others. This is close to the exact opposite of Mill's view. Of course, governments must make determinations of truth in the absence of unanimity among the citizenry. No government that both encouraged the free give-and-take of opinion and required unanimity on truth determinations before acting would be able to do anything. It is because governments *must* make truth determinations and act on them that the free give-and-take of opinion is so important. It is what makes the government's truth determinations rational, and thus reliable. Mill put it this way: "There is the greatest difference between presuming an opinion to be true, because, with every opportunity for contesting it, it has not been refuted, and assuming its truth for the purpose of not permitting its refutation" ([1859], 26).

Mill's position is that both individuals and governments must act on the basis of their beliefs. Success in achieving their goals depends on how reliable their propositional beliefs are, where those beliefs include not only beliefs about how to achieve their ends, but also beliefs about what is good for human beings and what is not. So both individuals and governments depend on the free give-and-take of opinion to be able to reliably achieve their goals.

The Millian Argument Does Not Depend on Distinguishing between Reasonable and Unreasonable Opinions

Although Mill thought that some opinions were reasonable and some were unreasonable, the distinction between reasonable and unreasonable opinions plays no role in the Millian argument. The argument is an argument for

freedom of all sincere expression of propositional opinion, whether reasonable or not. Of course, the view is that the free give-and-take of opinion tends to favor the reasonable ones over time, but this is only a tendency, not an exceptionless rule.

Because the Millian argument makes no distinction between reasonable and unreasonable opinions, it provides a marked contrast to much Anglo-American moral and political philosophy, especially the philosophy of human rights. Due largely to Rawls's influence, the reasonable/unreasonable distinction plays a crucial role in all these areas of philosophy. For example, in the philosophy of human rights, it is routinely taken for granted that something cannot be a human right if there is reasonable disagreement about it or that human rights must be part of an overlapping consensus of reasonable views (e.g., Donnelly 2003; Reidy 2008; C. Taylor 1999; von Platz 2008). It seems to me that this use of the reasonable/unreasonable distinction is a serious mistake. The distinction simply cannot bear the weight that has been placed on it.

How did this distinction come to play such a large role in moral and political philosophy, including the philosophy of human rights? The problematic use of the distinction can be traced to Rawls's use of the terms in *Political Liberalism*. As I discussed in chapter 7, in *Political Liberalism*, Rawls introduced the term *reasonable* and used it in two related senses. In the first sense, it applies to people who are willing to cooperate on fair terms of social cooperation—that is, those whose comprehensive views are part of an overlapping consensus on a liberal conception of justice for the political sphere (1993, 48–54). In the second sense, *reasonable* applies to those people's comprehensive moral, political, philosophical, and religious views—that is, those views that "recognize the essentials of a liberal democratic regime" (1993, 87).[4] In this second sense, Rawls believes that it is just a fact that there will always be *reasonable pluralism*—that is, disagreement among reasonable comprehensive views that overlap on the liberal conception of political justice.

The problem with using *reasonable* in this way is that there is a potential for confusion with its ordinary meaning, according to which a view is reasonable if there is some good reason for holding it and unreasonable if there is no good reason for holding it. In order to avoid confusion with the ordinary use of the term, I am going to adopt the convention I suggested earlier and use *reasonable*$_L$ for the special conception of reasonableness that Rawls employs in *Political Liberalism*. Rawls specifically refers to "views that reject one or more of the democratic freedoms" as unreasonable$_L$ (2003, 64, n.19).[5]

Note that using the subscript avoids our inferring, what would otherwise seem to be a straightforward consequence: that the Rawls of *Political Liberalism* is committed to thinking that any comprehensive view that denies any of the democratic liberties of political liberalism is unreasonable, in the sense that there is no good reason for opposing any of the democratic freedoms in political liberalism.[6] This would imply that there was nothing reasonable to be said on behalf of liberalism's two main opponents: libertarianism (e.g., Nozick 1974) and communitarianism (e.g., MacIntyre 1988).[7] It would also

dismiss as unreasonable the advocates of rule by a philosopher-autocrat (Plato's *Republic*), monarchists, anarchists, socialists, meritocrats, and moral skeptics. Can all of these views be dismissed as unreasonable, on any plausible notion of *unreasonable*?[8] Such a claim would be a return to the days of the Proof Paradigm, when philosophers saw themselves as engaged in proofs and, therefore, were committed to holding that no one could reasonably (in the ordinary sense) disagree with their views. Rawls would surely not want to make such an (epistemically) immodest claim.

Using the subscript prevents this misunderstanding. With the subscript, the claim is a trivial one: Any comprehensive view that denies any of the democratic liberties of political liberalism is unreasonable$_L$—that is, not willing to cooperate on the fair terms of social cooperation as agreed to in the original position and specified by the principles of political liberalism. Of course, libertarians are not willing to accept the principles of political liberalism.

The potential for confusion between ordinary reasonableness and reasonableness$_L$ was compounded when Rawls wrote *The Law of Peoples*. In *The Law of Peoples* he considered the question of whether liberal peoples have a duty to tolerate and cooperate with some nonliberal peoples, where tolerating includes respecting them, not just putting up with them (1999, 59–62). In *The Law of Peoples*, Rawls introduces a new standard of *decency* and explains it as a weaker standard than the reasonableness$_L$ standard (1999, 67). Thus he makes it quite clear that the rights that must be guaranteed for decency are a proper subset of the democratic freedoms of a liberal society. In fact, they don't even include democratic rights at all (65). Here is what Rawls says: "I am not saying that a decent, hierarchical society is as reasonable and just as a liberal society" (84). He says they "deserve respect" even if they are not "sufficiently reasonable from the point of view of political liberalism or liberalism generally" (84). And finally, of decent, nonliberal doctrines he says, "I do not say that they are reasonable, but rather that they are not fully unreasonable; one should allow, I think, a space between the fully unreasonable and the fully reasonable" (74). So now we find ourselves not merely with a two-part reasonable$_L$/unreasonable$_L$ distinction, but with a three- or four-part distinction: for example, fully reasonable$_L$; reasonable$_L$ but not fully so; unreasonable$_L$ but not fully unreasonable$_L$; fully unreasonable$_L$.

In addition, in *The Law of Peoples* itself, Rawls *extends* the liberal conception of reasonableness (reasonableness$_L$) to a conception of reasonableness that requires liberal justice only domestically and permits decent justice internationally. I refer to this conception of reasonableness as *reasonableness*$_{L+D}$. It is on this new conception of reasonableness$_{L+D}$ that Rawls can say that a Society of Peoples composed of liberal peoples and decent peoples is *reasonable* (1999, 5, 64, 68, 84) and *reasonably just* (5, 11) and that a Law of Peoples for both liberal and decent peoples is *reasonable* and just (83). And some of his uses of *reasonable* don't even seem to fit any of these categories, as, for example, when he says, "A people sincerely affirming a nonliberal idea of justice may still *reasonably* think its society should be treated equally

in a *reasonably* just Law of Peoples" (70, emphasis added). How can a nonliberal people be *reasonable*, if reasonableness requires a liberal conception of justice for domestic institutions? My best interpretation of what Rawls is saying here is that he is thinking that a liberal people could reasonably$_{L+D}$ think that the nonliberal people should be treated equally and the nonliberal people could endorse the liberal people's opinion.

In any case, Rawls's discussion in *The Law of Peoples* opens the door to even more special conceptions of reasonableness. As Rawls was well aware when he wrote *The Law of Peoples*, many liberals would disagree with him and insist on a liberal conception of justice internationally as well as domestically. They would be advocates of *reasonableness*$_{L+L}$. Then libertarians could present their views as a defense of *reasonableness*$_{LIB}$, communitarians as a defense of *reasonableness*$_{C}$, and so forth. There is no limit to the different special conceptions of reasonableness.

If even Rawls multiplies reasonableness distinctions, it was inevitable that when those distinctions were taken up in the political philosophy and human rights literature, things would get even more confusing.[9] So, for example, when Rawls says that liberal rights are rights on which there is no *reasonable*$_{L}$ disagreement, he invites the misunderstanding that he is asserting that there could be no reasonable disagreement on liberal rights, in the ordinary sense of *reasonable*. It is probably just this potential for misunderstanding that led him to extend his special conception of reasonableness in *The Law of Peoples* to *reasonable*$_{L+D}$, so that he could characterize human rights as the rights on which there is no *reasonable*$_{L+D}$ disagreement. But this only added to the confusion, because, in the ordinary sense, it is quite clear that there is lots of reasonable disagreement about his surprisingly short list of human rights.[10]

On the other hand, if we understand *reasonable* in the ordinary sense, then defining human rights as the rights on which there is no reasonable disagreement is disastrous for the philosophy of human rights. It is unlikely that there are any rights on which there is not *any* reasonable disagreement. Moreover, in the ordinary sense of *reasonable*, there is lots of reasonable disagreement on whether women should have any rights at all. This is bad news only if one thinks that philosophy progresses by stopping reasonable disagreement. Mill's message is that the opposite is true.

The result of Rawls's introduction of a no-reasonable-disagreement test for human rights has been that everyone who adopts the test must argue that there could be no reasonable disagreement with the rights on their list, even though they disagree among themselves on which rights pass the test. This disagreement could be due to the fact that they employ different special conceptions of reasonableness. But then the obvious problem is how the justification of rights as reasonable in a special sense of reasonable could possibly justify them to someone who reasonably disagreed with that special conception of reasonableness. And if the disagreement among human rights theorists on which rights pass the test is not due to differences in their special conceptions of reasonableness, then how could they possibly think that the rights on their list could

not be the subject of reasonable disagreement, when others who share their conception of reasonableness do reasonably disagree with them?

Perhaps we should ask those who use such distinctions to place the following disclaimer prominently in their works: "The special reasonable/unreasonable distinction that I employ in this book is not the ordinary distinction that you are familiar with. In the ordinary sense of the terms, there will be lots of reasonable disagreement with what I say in this book, including with how I draw my special reasonable/unreasonable distinction." I don't use the reasonable/unreasonable distinction to do any theoretical work in this book, but I want to take this opportunity to issue a similar disclaimer: Practically everything that I say in this book is subject to reasonable disagreement, including my views on which rights should be universally guaranteed as human rights.[11]

There is a deep problem with the whole idea of using the existence of reasonable disagreement as a test for whether something is a human right. The problem is that even the concept of reasonable disagreement is one on which there is and always will be reasonable disagreement.[12] Given this fact, why would anyone want to make the absence of reasonable disagreement a test of human rights?

In any case, it is quite clear that such a conception has no useful role in the history of human rights. An opponent of slavery in the eighteenth or nineteenth century could hardly have thought that there was a human right against slavery, if human rights could not be subject to reasonable disagreement. It is not even clear that we could have a coherent idea of human rights (as opposed to men's rights) if the concept required the absence of reasonable disagreement, because there is lots of disagreement today over whether women should have anything close to the same rights as men. Do we really want to insist that those who, on the basis of their religious beliefs, deny women most of the rights on the U.N. Universal Declaration of Human Rights are unreasonable? It is possible to think that there is room for improvement in their understanding of the role of women without insisting that they are unreasonable. And even if you think their religious views are unreasonable, what about the views of those who advocate strict divisions between the sexes on evolutionary grounds? Must we think they are unreasonable in order to be able to believe in equal human rights for men and women?

In the history of the development of every human right, there has always been lots of reasonable disagreement about it. Why think that a right is not a human right until all the reasonable disagreement has been resolved? I think the answer is connected to another question: How can we be justified in forcibly intervening to impose human rights on a society that does not recognize them if there is reasonable disagreement about them?

I think the criterion of no reasonable disagreement on human rights seems compelling because of an assumption that failure to guarantee human rights is a ground for forcible intervention. I think it is a mistake to think of human rights in this way. It should be possible to believe in human rights without thinking that it

is a good idea to try to impose them on others by force. Recall that the main principle evaluates practices both as substantive practices and as implementation practices. It is conceivable that the main principle would endorse human rights only when combined with noncoercive implementation practices.[13]

My impression is that there is another reason that motivated Rawls to look for rights on which there could be no reasonable disagreement. I think the longing for a position that is immune to reasonable disagreement is yet one more manifestation of the influence of the Proof Paradigm in Western philosophy. According to the Proof Paradigm, one could be justified in believing only something that was self-evident or provable from self-evident premises. Thus, according to the Proof Paradigm, reasonable disagreement was impossible. Even though the early Rawls explicitly rejected the Proof Paradigm in favor of an equilibrium model of reasoning, he never gave up the goal of articulating a free-standing political philosophy that would not be subject to reasonable disagreement. Thus, for all of his efforts, I wonder if Rawls ever freed himself from the influence of the Proof Paradigm.

In seeking a way to find a place for his philosophy above the fray of reasonable disagreement, Rawls was chasing a chimera. It was inevitable that there would be reasonable disagreement about Rawls's list of liberal rights and about his list of human rights, for there will always be reasonable disagreement about any interesting philosophical position.

Because of the contemporary identification of liberalism with a conception of liberal neutrality or of public reasons, it is often taken for granted that a distinction between reasonable and unreasonable disagreement is essential to liberalism—for example, to explain the difference between the concepts of the good that liberalism is neutral among and those that it is not, or to explain the difference between reasons that qualify as public reasons and those that do not. Thus, it is refreshing to be reminded that more than 150 years ago J. S. Mill defended a form of liberalism that employed no such distinction. Mill could have insisted that only those with reasonable views had a right to participate in the free give-and-take of opinion, but he did not. He thought that the actual process depended on everyone's being free to contribute, regardless of how reasonable or unreasonable their views were.

Mill never would have suggested that his epistemological argument for a right to freedom of thought and expression or that any of his arguments for autonomy rights were not subject to reasonable disagreement. He intended to be contributing to the free give-and-take of opinion, not ending it.

The Millian Epistemological Argument Does Not Undermine Itself

Because Mill's own theory of freedom of expression is itself intended as a contribution to the free give-and-take of opinion, it might seem that it undermines itself. After all, at the time that Mill proposed his epistemological

argument for freedom of expression, it was widely criticized, and even today it is unlikely that a majority of scholars in the field would accept it. By his own standards, then, hasn't Mill's theory been shown to be irrational by the process of the free give-and-take of opinion?

I don't know of any place where Mill directly responds to this question. I will give my answer based on how I would extend Mill's view, because the same challenge applies to my view. There are two strands to this question that need to be separated. First, the admission of fallibility (epistemic modesty). Mill does explicitly base his epistemology on human fallibility, so, of course, he would acknowledge his own fallibility. On philosophical issues especially, he would insist that it is extremely unlikely that any position contains the whole truth or that any sincerely held position has no truth to it at all. The free give-and-take of opinion makes possible improvements over time. In working out a philosophical position and then publishing it, we should be understood as exerting some influence on the process to nudge it closer to the truth. We present it not as the final truth, but as our best stab at getting closer to it.

It might seem that if the process is our way of getting closer to the truth, we should all just replace what we currently believe with the results of the process. Note that this is what we tend to do on subjects that we know nothing about. In such cases, often the quickest way to find out what we want to know is to do a search on Google or Wikipedia. But if we were to do this on subjects we know something about, the results would be disastrous. Everyone would have the same opinions about everything, and the process of the free give-and-take of opinion would grind to a halt.[14]

The process of free give-and-take of opinion works because individuals have access to or are attuned to different evidence or have different ways of thinking about the relevant issues or have different influences affecting them. This variety of different points of view is crucial for the success of the process. However, as participants in the process, we can recognize that the results of the process are generally more reliable than the opinions of anyone involved in the process (Surowiecki 2004). So the Millian argument does not undermine itself, because it explains why its adherents should advocate it. However, it does require them to do so with a certain modesty, because they will realize that it is almost certain that there are important objections to their views that they have not adequately addressed and that it is always possible that they have made a really big mistake without realizing it.

Another Paradox Resolved

In *On Liberty*, Mill imagines an interlocutor who says, roughly, *You are willing to use coercive laws to implement your philosophy of protecting freedom of expression. How can you object to my using coercive laws to implement my philosophy of censorship*? Mill's answer is that implementing

freedom of expression makes it possible to correct ourselves, even our views about freedom of expression. This is the right answer, but Mill did not pursue it far enough.

The interlocutor could rightly object that Mill has just upped the ante. The interlocutor says, *I started out challenging your basing coercive laws on your philosophy and you answered me with more of your philosophy. So we are still in the symmetric position that I described. You think it is OK to base coercive laws on your philosophy, but you deny me the right to base coercive laws on my philosophy*.

I am not sure how Mill would have responded to this elaboration of the objection, so let me say how I think he should have replied: Throughout most of history individuals have thought that if they could obtain political power they would be justified in using force to impose their moral and religious and philosophical views on others. In holding that political power should not be used this way, I am not attempting to articulate a view that is neutral among all views (not even all reasonable ones) about the use of political power; I am attempting to contribute to our only way of identifying and correcting past mistakes, the process of the free give-and-take of opinion. I am attempting to influence that process to generate a greater appreciation of the dangers of using political power to silence opinion.

And I would add the following: Just as we have discovered that to make progress in knowledge requires a process of free give-and-take of opinion, so we have also discovered that to make improvements in our government and laws, we should replace hereditary autocrats with a democratic process, constrained by human rights, that is itself based on the free give-and-take of opinion.

Mill's interlocutor thought that Mill's acknowledgment of his own fallibility implied that he was committed to thinking that all political views are equally good or equally worthy of being implemented. This is the mistake that I mentioned at the beginning of the previous chapter that has been made by almost all process epistemologies. They are all relativist. Mill's epistemology was fallibilist (epistemically modest), *not* relativist (metaphysically modest). In contemporary political philosophy, Habermas came the closest to articulating a view of this kind, but even he chose relativism rather than metaphysical immodesty. I choose metaphysical immodesty.

The Millian Argument Does Not Claim that the Process of the Free Give-and-Take of Opinion Is Free of Bias or Perfectly Reliable

Near the end of *On Liberty*, Mill considers the case of expression aimed at persuading you to do something that society regards as harmful. Mill says of such a case, "Whatever it is permissible to do, it is permissible to advise to do. The question is doubtful, only when the instigator derives a personal

benefit from his advice . . . subsistence or pecuniary gain" ([1859], 111). This is the closest that Mill comes to addressing what would later become a huge industry, advertising. In his discussion, Mill made it clear that some regulation could be justified. I want to expand on this topic.

Imagine the free give-and-take of opinion as taking place in a large group in which everyone is assigned a fixed location. Everyone's voice has a limited volume, so each person's opinions can be directly communicated to only a few people. For their opinions to spread in the group, they have to persuade some of those in their immediate vicinity, who persuade others in their immediate vicinity, and eventually the opinion could spread to the entire group, although probably not without some modifications along the way. One day some members of the group obtain megaphones. With their megaphones, they can communicate their opinions directly to a larger immediate group. Then some people get microphones and speakers that permit them to address the entire group at once.

These amplifiers reflect the effects of the press and other media, including the Internet, on the free give-and-take of opinion. These effects would not be a concern if there were no potential for making money by using the media to influence public opinion, because in such a case, those who were wealthy would have things they would want to spend their money on other than trying to influence public opinion. However, the case is different in areas such as advertising, in which there is a great pecuniary interest in influencing public opinion. If there were fairly equal pecuniary interests on all sides—for example, if those with a pecuniary interest in persuading people that smoking is not harmful were balanced by those with a roughly equal pecuniary interest in persuading people that smoking is harmful—the overall system could continue to work as Mill envisioned it. In the smoking example, potential smokers would hear both sides of the issue and then make their own decision. But in the case of advertising for products that may be harmful, although there may competition with other voices pushing other brands, everyone with a significant pecuniary interest will want to minimize any evidence that the products are harmful. Thus, the Millian epistemological argument for a right to freedom of expression is compatible with truth-in-advertising laws. Indeed, as I discuss in chapter 9, it may require mandatory disclosure laws, in cases in which, for example, drug manufacturers would otherwise suppress evidence of adverse side effects.

So the Millian argument for freedom of expression does not support any presumption that there is a right to what I call *voice amplification*. Sellers of cigarettes should be free to express their opinion on the safety of cigarettes, but there is no reason to suppose that they have a right to *amplify* it. This is important not only in product advertising but in advertising political campaigns. In political campaigns, money amplifies the candidate's message. Campaign donors can obtain great pecuniary advantage from influencing the outcome of an election. The Millian argument for freedom of expression creates no presumption against limits on political contributions or against limits

on campaign spending on advertising, so long as the limits are not so low as to prevent a challenger from waging an effective campaign against an incumbent. In this context, the decisions of the U.S. Supreme Court (e.g., *Citizens United v. Federal Election Commission*),[15] which have struck down limits on private campaign expenditures in the name of freedom of expression, are at best naive and at worst complicit in a regime that encourages what can only be described as legalized corruption. I return to this topic in chapter 10.

Another potential source of bias to the process of free give-and-take of opinion is the consolidation of media ownership. If consolidation were to proceed so far that one person or a small group could control the content of TV, newspapers, and radio in a single geographic area, then there would be a great potential for biasing the free give-and-take of opinion. However, it is not size of media holdings per se that is a problem. What is crucial is that there is a competitive media market. In a competitive market, media companies, no matter how large, will produce what people want to read or watch or listen to. So, for example, large capitalist corporations will be happy to sell copies of Marx and Engels' *Communist Manifesto* and Abbie Hoffman's *Steal This Book* (titled in large letters on the cover).

Not all sources of bias in the process of free give-and-take of opinion are due to the fact that money can be used to amplify one's voice. Some biases are less visible. For example, there is no country in the world that does not teach its children a biased version of its own history. In the United States, until the 1970s, there was practically no awareness of the disaster that European colonization was for American natives. Today there is a holocaust memorial museum in Washington, D.C., to remember the holocaust in Germany, but no memorial to the extermination and near extermination of native American tribes that took place in the United States, an American holocaust (Stannard 1992). Children in the United States learn practically nothing about the labor movement in the United States, certainly nothing about what prompted that movement, the miserable conditions in which millions of factory workers and other worked and lived in the United States in the second half of the nineteenth century. When the Smithsonian's National Air and Space Museum mounted an exhibit of the Enola Gay, the bomber that dropped the first atomic bomb on Hiroshima, the inclusion of photos of the devastation and a plaque that simply asked the question whether the bombing was justified caused Congress to threaten to close down the exhibit and led to the resignation of the museum's director. If certain reasonable questions cannot even be asked, how could the free give-and-take of opinion have a chance?

Confronted with the way that bias influences the process of the free give-and-take of opinion, we may start to feel pessimistic and wonder whether the process really is generally reliable. It is certainly not perfectly reliable, as we have seen, but is it reliable enough to, well, rely on? A Gallup poll reported that a majority of U.S. scientists (55%) but only a small minority of the U.S. public (10%) believe that human beings evolved from other forms of life with

no involvement from God.[16] In other countries the difference is much less pronounced. But there will always be examples of this kind.[17]

Such examples show beyond question that the social process of the free give-and-take of opinions will never be epistemically ideal. There will always be room for improvement. But they do not undermine the claim that, over time, the general tendency of the process is toward improvement—that is, to increase knowledge. This seems hard to deny. For example, any reasonable comparison of beliefs held generally 200 years ago with beliefs held generally today would make evident the great amount of progress that has been made.

It is also important not to overstate the significance of examples in which popular beliefs diverge from scientific belief. Most laypeople defer to scientists for scientific beliefs, so on most scientific matters, popular opinion will endorse accepted scientific beliefs, even without knowing what they are. The phenomenon of deferring to others regarded as authorities makes the social process of the free give-and-take of opinion remarkably reliable, when the other development-of-judgment and exercise-of-judgment rights are guaranteed.

We have two striking confirmations of the way that democratic process of identifying authorities can produce reliable results: the search engine Google and the online encyclopedia Wikipedia. The Google search algorithm is very complex, but the main idea behind it is very simple. On any topic, it rates Web sites that mention that topic on the basis of how many other sites link to them, where the linking sites are also weighted by the number of sites linking to them, and so forth. Google's rankings are based on popularity. Anyone in the world can vote, simply by creating a Web page with links. Each person's votes simply indicate which other sites they think are worthwhile. Whoever would have thought that such a completely open voting process would be such a good indicator of relevance and of truth? Google, founded in 1998, had by 2008 become an indispensable part of life for millions.

Much the same is true of Wikipedia, an online multilanguage encyclopedia that anyone may edit. There are articles in over 250 languages. Founded in 2001, by 2008 its English edition had over 2 million articles. This alone makes it the closest thing to the results of the free give-and-take of opinion that we will ever see. Inevitably, there have been some cases of "editing" to introduce falsehoods or to eliminate unwelcome truths. However, when *Science* commissioned scientific experts to review 42 scientific articles in Wikipedia and the *Encyclopedia Britannica*, the experts found an average of 3 errors per article in the *Encyclopedia Britannica* and only 4 errors per article in Wikipedia (Giles 2005). I doubt that anyone would have predicted that Wikipedia would do so well in a comparison with *Britannica*. Of course, errors in Wikipedia can be corrected immediately. Google and Wikipedia are two striking examples of the epistemic power of aggregate opinions.[18]

The Wisdom of Crowds and Human Rights

For most of history, philosophers have been so struck by the defects in the individual judgment of most people that they could not believe that aggregating those individual judgments could be produce anything of value. In 2009, there were so many books on the irrationality of human judgment (e.g., Ariely 2008) that it seemed to be something of a miracle that a rights-respecting democracy could even survive, much less flourish. No matter how many times it has been confirmed, the discovery that some groups (as opposed to mobs) are often better judges than any individual in the group has repeatedly come as a surprise. The jury system was not introduced because it was believed that groups of impartial jurors could do a better job than individual fact finders.[19] However, when a poll of 574 federal judges asked them whether they would want a case in which they were a party tried before a judge or a jury, they favored juries by an 8-to-1 margin (Curriden 2000). In politics, prediction markets are superior to polls at predicting the outcomes of elections. Companies have found that prediction markets are the best way for them to predict their own future (Sunstein 2006b). This is the phenomenon that Surowiecki (2004) calls the "wisdom of crowds."

It is possible to overstate claims about the epistemic virtues of aggregation. No single individual or group is very good at political predictions. When Tetlock (2005) evaluated the reliability of expert political predictions, he found that their reliability was not much better than chance. The advantage of a rights-respecting democracy over other forms of government is that it has a feedback mechanism for eliminating policies that work and improving policies that don't work based on their actual effects, not their predicted effects.

Taken together, examples such as Wikipedia, Google, and other similar phenomena show that, when the autonomy rights are protected, the process of free give-and-take of opinion is more reliable than Mill could have ever dreamed.[20] In addition, the process of the free give-and-take of opinion itself leads to discoveries about sources of unreliability in the process. It is due to the free give-and-take of opinion that we have become aware of the biases in the way history is taught and the voice amplification biases that I discussed earlier. It is through the free give-and-take of opinion that strategies have emerged (e.g., truth-in-advertising and campaign finance reform) and will emerge in the future to improve the process. Indeed, there is no other way of discovering sources of bias and other inaccuracies in the free give-and-take of opinion than by the operation of that very process. The process works best when the greatest variety of voices and points of view are included.

There is one more kind of expression that needs to be discussed fully. However, before I can discuss it, I need to complete my consequentialist account of a right to freedom of expression by explaining how the Millian epistemological argument interfaces with the main principle.

The Main Principle and Rights to Freedom of Expression (Propositional and Nonpropositional)

At the end of chapter 2 of *On Liberty*, Mill had introduced and explained a revolutionary social epistemology, but he had not actually shown how his epistemology supported, on consequentialist grounds, a right to freedom of expression. One direct route would have been to argue that truth is of such great value that true beliefs are of incomparable value, in which case, he would have a straightforward argument for maximizing the number of true beliefs (while minimizing the number of false ones). Mill did not take this route.

Another direct route would have been to argue that the free give-and-take of opinion was necessary for autonomy and that autonomy was an essential of human well-being. This would imply that maximizing human well-being required doing whatever was necessary to make people autonomous. As I discussed in chapter 4, Mill did make this argument early in chapter 3 of *On Liberty*. But it was not persuasive.

The primary Millian argument for rights to freedom of expression connects autonomy rights to well-being more indirectly. The following is my version of the argument, which is a generalization of Mill's.[21]

Each of us in living our lives is an investigator into this question: What is the best life for me? Although at one time people thought that there were authorities who could answer this question for us, we have discovered that all the authorities can do is to provide us with the accumulated wisdom gained from past experiments. There are no authorities to tell us how we can improve on the status quo.

Most people's lives would be greatly diminished without music, literature, dance, and the other arts. In all of these areas, progress depends on experiments. Most of the experiments are failures, but those that succeed can enhance the lives of millions. Imagine what a difference it would make to your life if your favorite kinds of music did not exist, or, even worse, if there were no music at all.

Even if these activities do not aim at truth, what they aim at is promoted by the same process of free give-and-take of expression (not necessarily opinion). Most of the great advances in the arts have been opposed by those in authority. In Shakespeare's time, preachers ranted against the evils of the theater. Books of fiction, including what are now regarded as literary classics, have been burned throughout history. Salmon Rushdie was threatened with death for writing *The Satanic Verses*.[22] In my youth, preachers and other authorities ranted against rock music as the work of the devil; in 2009 there is an entire category of Christian rock music.

Most people may never experiment with making music themselves, but they benefit immeasurably from those who do. Although Mill argued that autonomy is an essential for well-being, this is far from true. Originators of great transformations in the arts can live miserable, tormented, foreshortened

lives plagued by drug addiction or alcoholism or mental illness. Whatever joy or other benefit Beethoven got from composing and conducting his Ninth Symphony is negligible in comparison to the joy and other benefits that others have derived from it.

The examples from the arts can be generalized. All of us are conducting experiments in living, in which the question to be answered is this: What is the best kind of life for me? Each of us benefits from the fact that others have conducted and are conducting parallel experiments with their own lives. Even someone who decides that a life of conformity is best for them is benefiting from all the past experiments that have influenced the current state of the society to which they want to conform.

This provides the basis for a generalization of the epistemological argument for freedom of opinion to an argument for freedom of expression, both propositional and nonpropositional. Literary and artistic expression makes at least as significant a contribution to human well-being as science and other disciplines that aim at truth. The free give-and-take of expression is the vehicle for progress not only in truth, but also in human well-being, because it encourages experiments in living and makes it possible for everyone to benefit from the experiments of others, both from those that are successes and from those that are failures.[23]

The Consequentialist Case for a Robust Autonomy Rights, Including a Right to Freedom of Expression

We are now in a position to put together all the parts of Mill's argument for freedom of expression. The social process of the free give-and-take of *opinion* is necessary for having rational beliefs that, over time, approach the *truth*. This makes possible a parallel social process in experiments in living that over time is the main source of progress in every aspect of human life. Successful experiments in living lead to progress, because successful experiments can be adopted by others. The linchpin of the entire process, the essential condition for the process to be progressive, is what I refer to as the *claim of first-person authority*: Given appropriate background conditions (specified by the autonomy rights, including freedom of expression), normal adult human beings are reliable judges of what is good for them, and generally more reliable than other people (especially government officials). This claim seemed absurd to many of Mill's [1859] contemporaries when he first made it. It is still controversial today. In my first volume (Talbott 2005, 123–127), I argued that the claim of first-person authority is true. In coming chapters I provide more supporting evidence.

The full Millian consequentialist defense of autonomy rights, including rights to freedom of expression (both propositional and nonpropositional), then is that experiments in living are necessary for progress in determining what kind of life is best for human beings, and thus for promoting life

prospects. Autonomy rights are the rights that increase both the probability of individuals' conducting successful experiments and the probability of other people incorporating the results of successful experiments into their own lives.

In my first volume, I added a second kind of consequentialist defense to the Millian defense of autonomy rights. Autonomy rights provide the necessary background conditions for governments to be able to make reliable determinations of which of their policies are successfully promoting well-being (and which are not) and for making governments appropriately responsive to that feedback, so that they act to equitably promote well-being. So there are two strands to the consequentialist case for a right to freedom of expression, as well as for autonomy rights generally.

Why *Robust* Autonomy Rights?

I have not addressed the question of why the main principle would support a robust right to freedom of expression. Why doesn't the main principle justify an exception whenever the government thinks that the exception will better promote (appropriately distributed) well-being or whenever a majority votes favors making an exception? This is my version of the question that, as I discussed at the beginning of this chapter, the early Rawls tried but failed to give a consequentialist answer to: What is the basis for the *priority* of the basic liberties over other legislation based on well-being considerations? My answer to this question has four parts:

(1) *The paradox of direct consequentialism*. In chapter 5, I showed how the paradox of direct consequentialism undermines any presumption that the main principle would justify the government in directly applying the main principle. Although this conclusion applies generally, it has particular force in the application to rights to freedom of expression. We know from the history of repression of expression that authorities are especially prone to misjudgment when they make judgments about the likely bad effects of political, moral, and religious opinion, as well as works of literature and of art. The main principle will not justify giving governments any authority to censor these opinions, except to avoid a "clear and present danger" of serious harm.

(2) *Minority versus majority rights*. Some rights—for example, consumer protection rights—are rights of majorities (consumers) against minorities (a business or corporation). Rights of majorities don't need to be legislated as robust rights. Autonomy rights, including rights to freedom of expression, are rights of minorities against majorities. They are rights that protect a nonconforming individual or group, even when a majority is upset and offended by the nonconformists. For these rights to be effective, they have to be robust enough to prevail over majority opinion.

Waldron (1998 and 2006) has argued that majorities should be the final arbiters of the content of rights, not courts. His argument is based on

democratic scruples. How could an individual judge or group of judges have the authority to overrule a majority? This seems to me to be a mistake. Majorities can and do oppress minorities. Other things being equal, the main principle supports governments that can protect minorities against oppression. In addition, as Mill argued, in the long run, tolerating nonconformists provides great benefits. Of course, Waldron would not accept this consequentialist justification. I discuss Waldron's position more fully when I take up the consequentialist justification of democracy in chapter 10.

(3) *The priority of truth*. The process of free give-and-take of opinion is our best method of making progress toward the truth. Mill's defense of this process does not require assigning any value to the discovery of truth per se. Pragmatists such as Stich (1990) have argued that truth is only one among many values, with the clear implication that it is not anywhere near the most important. But Kornblith (1993) has argued in response that there is a problem with subordinating truth to other values. The problem is this: Whatever values you have, you will need to determine the most effective ways of achieving those values. In doing that, you will depend on having true (or approximately true) beliefs. If your beliefs about how to achieve your values are largely false, you have very little chance of achieving your values.

This is an argument for giving priority to whatever is necessary for accurate determinations of relevant truths. Of course, that is the free give-and-take of opinion. To compromise the reliability of that process in order to pursue other values will ultimately tend to frustrate us in the pursuit of all of our other values. If governments are to equitably promote the life prospects of their citizens, they need the free give-and-take of opinion to perform two roles: first, to enable them to be able to obtain reliable information to use in the design and implementation of laws and other policies and, second, to assure that the feedback they receive from their citizens on how well their policies are promoting their well-being is reliable. Because this information or feedback is necessary for the success of any policy, the rights that are necessary to obtain it—that is, the autonomy rights—take priority (though not absolute priority) over any other policy.

(4) *The potential for abuse*. The greatest threat to the free give-and-take of opinion is the coercive power of government. Any exception to the right to freedom of expression has a potential for abuse—as a general rule, the larger the exception, the greater the potential for abuse. It is a truism that if a power can be abused, it will be. So any exception to the right to freedom of expression must be designed to prevent its being abused. This is the reason that exceptions to the right must be narrowly defined. Also, it is a reason for requiring those who will be tempted to abuse the exception to justify their actions to an impartial judge. Making the right to freedom of expression a robust right guarantees judicial review of government decisions to abridge it.

And, thus, for all four of these interconnected reasons, autonomy rights, including rights to freedom of expression, must be robust. They will also need to be inalienable, for reasons I explain in chapter 10.

A Comparison of the Millian Account with the Nonconsequentialist Accounts of Rawls and Habermas: The Toleration of Intolerant Subversive Advocacy

In the course of my discussion of the free give-and-take of opinion, I have indicated how the Millian argument would address most of the standard exceptions to freedom of expression. There is one standard case that I have waited until now to address, because it provides a good basis for comparing my consequentialist account with nonconsequentialist alternatives. The example involves freedom of political expression of a special kind, intolerant subversive advocacy. Subversive advocacy is the advocacy of the violent overthrow of the government. It is intolerant when in it advocates the establishment of a new government that does not permit free expression of unorthodox opinions. The question then is whether a liberal society should define the right to freedom of expression so broadly as to include the protection of intolerant subversive advocacy. This was the issue in an important U.S. Supreme Court case discussed by Rawls (1993), *Dennis v. United States*.[24] After briefly describing the issues in *Dennis* and the Court's decision, I consider how Rawls's political liberalism and Habermas's ideal discourse theory would apply to the case.

The *Dennis* case addressed the constitutionality of the Smith Act, a law passed in 1940 that outlawed subversive advocacy. It was used to prosecute members of the Communist Party in the United States. The Communists engaged in intolerant subversive advocacy, because they advocated replacing the U.S. government with a communist government that did not guarantee any right to disagree with communist party doctrine.

In *Dennis*, members of the communist party challenged the Smith Act on the grounds that the First Amendment right to freedom of expression protected their intolerant subversive advocacy, even though, were they to obtain power, they would not respect the rights of others to disagree with them. The standard applied by the court was the "clear and present danger" test. At the time, it was generally agreed that there was no imminent threat of violence. Indeed, the Communist Party membership had been declining for years. In applying the "clear and present danger" test, the court adopted Judge Learned Hand's statement of the rule, which measured the danger by taking the magnitude of the potential evil and discounting it by its probability.[25] Applying this test, the Court determined that even though the probability of overthrow of the government was very low, overthrow of the government was such a great evil that the Communist Party did represent a "clear and present danger" and, on that basis, upheld the constitutionality of the Smith Act.

However, this is no longer the Court's position. The current Supreme Court doctrine, enunciated in *Brandenburg v. Ohio*,[26] is that even intolerant subversive advocacy is to be tolerated unless it is likely to lead to imminent and unlawful use of force. Most commentators, including Rawls (1993, 345),

think that *Dennis* was wrongly decided and approve the *Brandenburg* result. The challenge for the nonconsequentialist accounts is to explain why *Dennis* was wrongly decided.

Rawls and Intolerant Subversive Advocacy

Because Rawls discusses at great length why the *Dennis* result was a mistake (1993, 340–356), it comes as something of a surprise to realize that what he says about *Dennis* is at odds with his own theory. This is such a surprising result that it will take me some time to set out all the pieces of the theory necessary to sustain my claim.

Rawls discusses subversive advocacy and the *Dennis* case in Lecture VIII of *Political Liberalism*. Earlier in that lecture, he had discussed liberty of conscience, as applied to religious, philosophical, and moral views. The arguments that he gave for liberty of conscience were all original position arguments—that is, arguments that were to be considered in the original position, behind the veil of ignorance. One of the surprises of Rawls's discussion of the toleration of subversive advocacy is that none of it makes use of the original position device. I believe that that is because an original position argument would *not* support Rawls's discussion of *Dennis*. To see why not, consider his original position argument concerning freedom of conscience. The argument shows that all parties in the original position would agree to tolerate those who reciprocated toleration. In Rawls's terms, reasonable$_{L}$ comprehensive views would all agree to tolerate other reasonable$_{L}$ comprehensive views. What about the unreasonable ones? Nothing that Rawls says in this argument answers that question. However, I think this is simply an oversight, because there is other textual evidence that makes it clear that the liberty of conscience agreed to in the original position would *not* cover unreasonable views.

The first evidence is Rawls's discussion of exactly the same issue in *Theory*. There he made exactly the kind of argument we would expect him to make in the original position, a reciprocity-based argument. In the original position everyone would be expected to agree to reciprocate tolerance. Rawls drew the explicit conclusion that the liberty of conscience that would be agreed to would only apply to the tolerant. There would be no duty to tolerate the intolerant (1993, 216–218).

Other textual evidence in *Political Liberalism* makes it quite clear that he had not changed his opinion on this issue. First of all, it is clear that those who claim the right to use state power to repress other reasonable (or not unreasonable) views are themselves unreasonable (1993, 61–62). What is the proper way for those who are reasonable to treat those who are unreasonable? Here is what he says about unreasonable doctrines: That there are such doctrines "is itself a permanent fact of life, or seems so. This gives us the practical task of containing them—like war and disease—so that they do not overturn political justice" (64 n. 19). Rawls does not specifically say what "containing" permits, but the comparison to war and disease suggests that

suppressing them—for example, by making them illegal—would be one permissible way of containing them. In any case, it is quite clear that liberty of conscience does not cover them.

The argument for liberty of conscience in *Theory* and Rawls's discussion of unreasonable comprehensive views in *Political Liberalism* make it clear that the basic liberty of conscience does not include unreasonable views, including intolerant subversive advocacy. What can we conclude from this? I think we have to conclude that at the constitutional stage, the stage at which the principles agreed to in the original position are translated into a constitution, the constitutional right to liberty of thought or conscience would extend only to reasonable views. The reason for this is simple. The constitutional rights give priority to the basic liberties, so that they cannot be traded off against considerations of well-being. At the constitutional stage, it would not be permissible to use well-being considerations to enlarge the basic liberties agreed to in the original position. It would be particularly inappropriate to include unreasonable views in a constitutional right to liberty of conscience when the appropriate attitude toward such views was to try to contain them as one would contain war or disease.

And now we can see why Rawls's discussion of *Dennis* made no use of the original position. The main considerations that Rawls uses to justify tolerating the intolerant are well-being considerations. The most prominent one is that a democracy can be made more stable by allowing the expression of subversive advocacy, even intolerant subversive advocacy, because, echoing Mill, those who advocate such views usually have some genuine legitimate grievances. Allowing the grievances to be expressed makes it possible for the government to address the legitimate grievances and thus remove some of the motivation for overthrowing the government (1993, 346–348).

Thus, we can see that, if we apply Rawls's own theory to the example, we must conclude as follows: First, there would be no constitutional right that required tolerating intolerant subversive advocacy. Second, if there were a case for tolerating intolerant subversive advocacy, it would strongly depend on well-being considerations, and thus would be within the purview of the legislature. Finally, whatever the legislature decided to do, whether to enact a law requiring tolerance for the intolerant or to enact a law making intolerant subversive advocacy illegal, would not be subject to judicial review, because the courts would have no constitutional basis for invalidating the law. It follows therefore that, in spite of what he says to the contrary, Rawls's own theory supports the *Dennis* decision. In a society based on *Political Liberalism*, Rawls would be free to try to influence the legislature to pass a law guaranteeing tolerance of intolerant subversive advocacy, but there would be no constitutional basis for the courts to overturn legislation, such as the Smith Act, which outlawed it.

On reflection, this result is not surprising. The original position is meant to capture a conception of moral reciprocity (Rawls 1971, 14). It is hard to imagine a greater departure from moral reciprocity than the position of

someone who says you should tolerate my view that advocates the forcible overthrow of your government even though after I overthrow your government, I will not tolerate your views.

The main problem with Rawls's theory is that it is an ideal theory. Because the principles of justice are chosen in the original position in which the conditions of choice assure that everyone will reciprocate cooperation, Rawls's theory does not have the resources to systematically address questions about nonideal theory, in which not everyone is not so cooperative. Intolerant subversive advocacy is an issue for nonideal theory.

It is to Rawls's credit that in *Political Liberalism* he presents a very strong argument for tolerating intolerant subversive advocacy. It is always possible to claim that the argument depends on some feature of the original position. But whatever role it might play, it seems clear that the decisive considerations have to do with (appropriately distributed) well-being.

Habermas and Intolerant Subversive Advocacy

The discussion of Habermas's theory can be briefer. Habermas never discusses the issue of tolerance of intolerant subversive advocacy. However, on Habermas's (1996) account, democracy gets it legitimacy from its being an approximation of rational discourse. Habermas's explanation of the role of constitutional rights is that they are to establish the conditions for democratic discourse to approximate the ideal of rational discourse. Thus, constitutional rights play the role of establishing the presuppositions of rational discourse.

Recall the four most important norms that govern the ideal speech situation: (1) publicity and inclusiveness—the discourse is open to all; (2) equal rights to engage in communication; (3) exclusion of deception and illusion; and (4) absence of coercion (2003, 106–107). Habermas's norms would exclude Communists from participation in ideal discourse, unless the Communists gave up their intolerant position and allowed everyone an equal right to engage in communication. The presuppositions of discourse would eliminate the intolerant at the outset. All of the participants would have to be tolerant.

I suppose it is logically possible that ideal discourse among tolerant participants would lead to an agreement that toleration for intolerant opinions would be equally in everyone's interests, but it is hard to see how it could ever come about. In theory, Habermas's construction needs to have someone to represent the position of the intolerant. But someone who held that position would fail to satisfy the norms of ideal discourse, especially equal rights to engage in communication, and so could not participate.

Consider a related case. Suppose that Osama believes that women should not be allowed to speak when men are present. If the issue is to be taken up by the process of ideal inquiry, it is hard to see how Osama could participate, because he will refuse to listen to the opinions of the women in the group.

Indeed, he will either try to suppress their expressions of their opinion or walk out himself. If he tries to suppress their opinions, presumably he would not be allowed to participate. But if Osama is not allowed to participate or if he walks out, who will argue for the right to espouse the view that women should not be allowed to speak when men are present?

Well, let's suppose that somehow the issue comes up for discussion. How in the world could the participants agree that it is equally in everyone's interests to allow the intolerant to express their intolerance, when the intolerant won't reciprocate? It seems to me that, like Rawls, Habermas has described an ideal process that would yield a moral ideal of reciprocity. Neither ideal process can explain why the right to freedom of expression appropriate for our nonideal world does not require reciprocity. In an ideal world, no one would have intolerant opinions, so the question of whether to tolerate the intolerant would never come up. In our nonideal world, the working of the social process of the free give-and-take of opinion depends on freedom of expression of all opinions, reasonable and unreasonable, because allowing governments to censor opinions they deem unreasonable is a social practice with very bad consequences.

Because Habermas's ideal discourse would not support a right to freedom of expression that included intolerant subversive advocacy, Habermas's theory cannot explain why the *Dennis* decision was a mistake either. Like Rawls, Habermas offers an ideal theory. It really is true that in an ideal world, no one would claim a right to freedom to express the opinion that other people should not be free to express their opinions. So an ideal theory would not even consider intolerant subversive advocacy. The main principle applies to our nonideal world.

Other Nonconsequentialist Theories

Although Rawls's and Habermas's nonconsequentialist theories do not imply that *Dennis* was decided wrongly, some nonconsequentialist theories, such as libertarianism, give the right result in this case, but not for the right reasons. Libertarianism, for example, implies that almost all laws are invalid, including many quite legitimate ones. Even if a libertarian state is not as minimal as Nozick's (1974) minimal state or the night watchman state of classical liberal theory, there are lots of things that a libertarian state could never do that the main principle would easily endorse, most prominently, providing legal solutions to collective action problems, as I discuss in chapter 10.

Intolerant Subversive Advocacy and the Main Principle

The application of the main principle to intolerant subversive advocacy is straightforward. Indeed, Rawls's extended discussion of the topic in *Political Liberalism* fits the main principle better than it fits his own theory. The

main principle gives priority to rights that are necessary for the free give-and-take of opinion. For the reasons that Rawls discusses, intolerant subversive advocacy makes an important contribution to the free give-and-take of opinion. It enables governments to discover justified grievances and to resolve them peacefully. As a result it makes governments more stable, rather than less. Of course, the main principle always allows for exceptions. Surely, there would be an exception for cases in which there was a genuine "clear and present danger" of violent insurrection, but this test could not be interpreted using the Hand formula as the *Dennis* court did. It would only apply in the kind of situation that Rawls refers to as "a constitutional crisis of the requisite kind" (1993, 354). I agree with him that the United States has probably never had such a crisis. What is ironic is that allowing freedom of expression, including intolerant subversive advocacy, which would make an autocracy less stable, typically makes a rights-respecting democracy more stable.

The Other Autonomy Rights

In this chapter, I have focused on the right to freedom of expression, but the discussion can easily be applied to the other exercise-of-judgment rights—that is, rights to freedom of assembly and freedom of the press. Due to the work of Sen (1999, 178–188), it is now appreciated that freedom of expression and freedom of the press play an essential role in combination with democratic rights in eliminating famines and other disasters. A free press is essential to preventing and correcting a great variety of government abuses. It is so important to the functioning of a democracy that no system of government that lacks a free press should be classified as a true democracy.

The main principle's endorsement of development-of-judgment rights—rights to physical subsistence, to what is necessary for normal development, and to education, in addition to security rights—is straightforward. In order to be able to exercise good judgment, you need to develop it. There is much more that could be said about the contours of a right to education, but it is too large a topic to undertake here.[27] Because the development-of-judgment rights and the exercise-of-judgment rights are essential parts of the solution to the reliable feedback problem and the appropriate sensitivity problem, they all have priority over other legislation, and thus they should all be regarded as robust rights (Talbott 2005).

Conclusion

We have seen that there is a two-part consequentialist argument for freedom of expression and freedom of the press that covers personal expression of opinions with propositional content (understood broadly, to include

normative and evaluative opinions), as well as literary and artistic expression with and without propositional content. The first part of the argument is epistemological: These rights are necessary for the free give-and-take of expression. The second part of the argument is consequentialist: The free give-and-take of expression is necessary for progress in equitably promoting well-being.

NINE

Property Rights, Contract Rights, and Other Economic Rights

> Then will not lawsuits and accusations against one another vanish, one may say, from among them, because they have nothing in private possession but their bodies, but all else in common? So that we can count on their being free from the dissensions that arise among men from the possession of property . . .
>
> —Plato

> I am persuaded, that till property be taken away, there can be no just or equitable distribution made of things, nor can the world be happily governed.
>
> —Thomas More

> Property is theft.
>
> —Proudhon

> The theory of the Communists may be summed up in the single sentence: Abolition of private property.
>
> —Marx and Engels

> I'm a marketable commodity.
>
> —Cher

Economic rights—for example, rights to property, markets, and contracts—often are overlooked or given short shrift in discussions of human rights, perhaps because human rights are typically aimed at ending oppression and economic rights have often been used to justify it. There is also a problem of perception. The negative effects of economic rights are obvious: egoism, greed, and inequalities in wealth. Their positive effects are much less obvious, though more profound. Finally, economic rights are sometimes defined in terms of markets free from government regulation, when, on the contrary, it is appropriate government regulation that makes markets worth having. In spite of this perception problem, when combined with the other items on the list, economic rights are important human rights.

What Are Property Rights?

The main division in theories of property rights is the division between those who regard the principles of ownership as inherent in the nature of things and thus immutable and those who regard property rights as social constructions, what Rousseau called "conventions."[1] Those who regard them as social constructions may disagree on what their point is—for example, to promote autonomy or nondomination or efficiency or overall happiness or, as I would hold, to equitably promote well-being. Hypothetical consent theories of property rights are also social constructionist, because, even if they deny that there is anything that property rights are aimed at promoting, they provide a criterion on which any existing system of property rights principles could be improved.

Immutabilists deny that human beings have the power to alter the inherent nature of ownership. They include those who think it is an immutable truth that there is no such thing as ownership. They also include the various libertarian views of the nature of property. Somewhat surprisingly, they also include Marxist views, according to which labor necessarily creates ownership of what it produces (G. A. Cohen 1995).

Common sense favors the immutabilists. It does seem to be an immutable truth that if I go prospecting for gold and find it on land that no one else has a claim to, then I am entitled to claim it for my own, or that if I find a plot of land to which no one else has a claim and till and plant and cultivate an apple orchard, I am morally entitled to ownership of the apples it produces and that other people who contributed nothing to the production of those apples would be prohibited from eating those apples without my permission.[2] It will be useful to have names for these two ways of acquiring ownership, the first, by *discovery* of unowned things, and the second, by *mixing one's labor* with unowned things.

If the social constructionists are correct, this impression of immutable principles of ownership is a powerful illusion. To draw attention to the power of the illusion, Murphy and Nagel (2002) refer to the "myth" of ownership. It is important to be clear on what it is that they are claiming is a myth. On their view, it is not ownership itself that is a myth, it is the view that the principles of ownership are immutable principles in the nature of things. Property exists, but the principles of ownership, the institution or social practices of ownership, is a social construction.

Some sort of ownership principles are almost surely hardwired into our cognition, because we find what might be thought of as practices of proto-ownership in many different species. Territoriality is a kind of proto-ownership practice.

If we are hardwired in a way that makes the principles of ownership seem immutable, there is no easy way to decide the issue between the immutabilists and the social constructionists. Fortunately, there is a category of ownership relations that evolution has not hardwired into us, so it can be used as

a test case to help us to decide between the immutabilists and the social constructionists (as well as to decide among the variety of social constructionist views). This is the category of intangible property—for example, copyrights and patents. Intangible property provides an interesting test case to help to decide between immutabilist and various social constructionist views.

Intangible Property

The most important categories of intangible property are copyright (ownership rights to an original creative work, but not allowing for ownership of scientific or mathematical laws), patent (ownership rights to a new invention—for example, a formula or process), and trademark (ownership rights to an identifying name or symbol). These three kinds of ownership rights are typically of different duration. In the United States, copyrights expire 70 years after the author's death; patent rights expire after 14 or 20 years depending on the nature of the invention; and trademarks can be renewed in perpetuity.

Intangible property rights pose a challenge for the immutabilist, because it is difficult to see how expiration dates for property rights could be part of the nature of things. Thus, we would expect an immutabilist to occupy one of the extremes, represented by the discovery and the mixing of labor models of original ownership: (1) The discovery model supports perpetual ownership—the person who discovers an invention or authors a creative work or trademark would acquire perpetual ownership rights (which, of course, can be transferred to others); (2) the mixing one's labor model supports no ownership in abstractions—an author's or inventor's ownership rights would extend only to the original manuscript of a work, because that would be the physical object with which the author or inventor mixed his labor. If the author or inventor made a copy of his work and sold it, he would have no rights to other copies made from that copy, because he would not have mixed his labor with them. Of course, an author or inventor could sell a copy of his work with a provision that prohibited the buyer from making copies without the owner's permission and made the buyer liable in damages for any copies made by others. But if someone else made a copy of the copy, the author or inventor would have no recourse against that third party, who, of course, would not have been a party to the original contract of sale.

The interesting thing about these two alternatives is that we can recognize that both of them represent a way of extending the idea of original ownership in tangible property to intangible property and it is hard to see how, *a priori*, one could decide between them. In addition, it is hard to see how anyone could really want to. To choose perpetual ownership, we would have to be willing to accept tens and probably hundreds of thousands of additional deaths each year and the number would increase over time, because, even with the help of relief agencies, most people in the world would not be able afford to buy medicines that were covered by perpetual patents. It is true that

tens of thousands have already died, because patents don't expire for 20 years. But there has also been a substantial benefit from patents. With the help of relief agencies, at any particular time almost everyone in the world can afford many of the medicines that constituted the highest standard of care 20 years earlier.[3]

On the other hand, if there were no intangible property rights, there would be no incentive for pharmaceutical companies to develop new medicines. On this alternative, everyone in the world would be worse off.

Aware that neither of these two options was an attractive one, Nozick (1974) made a suggestion that has been picked up and developed by A. Moore (1997). Nozick suggested that perhaps expiration date of intellectual property rights represents the average interval of time between first and second discovery (1974, 182). This proposal has the advantage of providing some sort of rationale for an expiration date for rights to intangible property, but it does so in a way that threatens to undermine their theory of rights to tangible property, for several reasons.

First, libertarian theory is a historical theory. Property rights depend on what actually happened, not on what would have happened. No libertarian theory of tangible property would hold that a prospector's ownership to a gold mine should be limited by the expected amount of time it would have taken for another prospector to discover it. Suppose there were other prospectors in the area who almost surely would have discovered it within days of the initial discovery. No libertarian would think that the fact that someone else would have discovered the mine had any effect on the duration of the property rights of the prospector who actually discovered it.

Second, because libertarian theory is a historical theory, *average* time of second discovery would seem to be irrelevant to a case in which we know the time to second discovery. For example, in science it often happens that when two labs are in competition, one lab makes the discovery only weeks or months before the other. When we *know* the time to second discovery, why wouldn't that fact limit the duration of the intangible property right rather than an average?

Third, it is just not plausible that the expiration period of copyrights and patents is any kind of an estimate of the time to second discovery. No one thinks that someone would probably have written *War and Peace* within 70 years of Tolstoy's death if he had not written it. There are no legal ownership rights to mathematical theorems and proofs or to laws of nature, though it is clear that in many cases (e.g., the discovery of Newton's laws) it could have taken years if not decades before they were independently discovered.

The only remotely plausible explanation for the limited duration of intangible property rights is that they represent a balance between two conflicting factors, the need to give authors a long enough period of time to able to benefit from their creations and inventions to motivate them to create or invent them and the great potential for public benefit from ending ownership rights and placing their creations and inventions in the public domain.

It is important to distinguish between the fact of creation or invention and the ownership rights that enable a creator or inventor to benefit from a creation or invention. Creation or invention is not typically a social construct. Tolstoy will be the author of *War and Peace* in perpetuity. Intangible property rights are rights to benefit from a creation or invention. No immutabilist has proposed any plausible explanation of why those kinds of rights would expire, if ownership rights in tangible property do not.[4]

Perhaps the immutabilist should adopt a hybrid view, immutabilism about rights to tangible property and social constructionism about rights to intangible property. But once the camel of social constructionism gets its nose into the tent of property rights, it is almost inevitable that it will try to get the rest of its body in too. The same kinds of considerations that explain the contours of rights to intangible property can also explain why many of the changes to property and contract law that have taken place over the last 200 years are improvements. That is the topic I turn to next. Perhaps surprisingly, once we understand the consequentialist rationale for the improvements to property and contract rights defined by a suitably regulated system of markets and civil liability, we are in a position to understand why those rights should be recognized as universal human rights.

The Historical Importance of Economic Rights in the Development of Human Rights

Although economic rights are not basic human rights, they play an extremely important role in the historical development of human rights and in my theory of human rights. I suspect that they were essential to the historical development of human rights, for a number of reasons. First, they gave monarchs an incentive to limit their own power, something that most of them would not have been inclined to do otherwise. The reason is very simple. Monarchs discovered that by protecting private property rights, they could generate wealth that could be taxed. Thus, it was in their interest to protect private property rights (Olson, 2000). Had it not been for the fact that private property rights generate wealth, human history might have continued the trajectory that is quite visible through the heyday of the European, Islamic, and Chinese empires, a history of the evolution of ever more oppressive forms of government exercising ever more oppressive techniques of social control.

Second, the generation of wealth created multiple centers of economic power that could, to some extent, counterbalance the power of monarchs. It is no accident that the development of political rights in Great Britain and in the United States was closely tied to issues of taxation. Had monarchs not needed to maintain a healthy economy, they could have and almost surely would have simply crushed the early movements for political rights.

Third, market economies could never have been successful unless people were able, at least in economic matters, to develop good judgment—that is,

the ability to be reliable judges of what is good for them. Market exchanges generate wealth because the exchanges typically make both parties better off.[5] If the parties were not reliable judges of what would make them better off, market economies would not generate wealth and they would eventually disappear. Because autonomy rights are the rights necessary for the development and exercise of good judgment, market economies and autonomy rights go together (e.g., Friedman 1962).

Economic Rights as Solutions to the Productive Investment Collective Action Problem (CAP)

Though no one predicted it *a priori*, in retrospect we can see that economic rights solve an important problem. Hobbes [1651], for example, argued that without a government to enforce property rights, everyone would claim ownership of everything, with the result that there would be no industry and life would be nasty, brutish, and short.[6] Let's call this the *productive investment CAP*: the problem of motivating people to productively invest their time to make improvements in the world. If there are no secure property rights, productive investments will be pointless because there is no way to prevent the product from being appropriated by others.

Of course, private property rights are not the only response to the productive investment CAP. Indeed, as the epigraphs to this chapter remind us, for almost all of human history, many intelligent people have thought it was obvious that private property rights were not a very good response to the productive investment CAP. Market economies did not come into existence because smart people predicted that they would successfully solve the productive investment CAP. They came into existence because people were motivated to make exchanges and, by a process of trial and error over thousands of years, changes that facilitated exchanges tended to motivate productive investments.

Today, market economies have been so successful in promoting productive investment that it is only a slight exaggeration to say that there are two kinds of successful governments, those that have market economies and those that profit from market economies by selling natural resources to them. Marx made one of the defining marks of Communism the abolition of private property. Communism died as an ideology when the Chinese government, though still nominally "Communist," reinstated private property rights in the 1980s and economic growth exploded.

If economic rights are favored as a solution to the productive investment CAP, it is important to point out that this is not something that could have been or was known *a priori*. It was because markets tended to reward productive investment that they tended to develop and grow. The discipline of economics did not come into existence until markets had already reached an advanced stage of development. When he wrote his great treatise on

economics, Adam Smith [1776] did not have to predict the economic benefits of markets and the division of labor; all he had to do was recognize them and explain them, which were significant accomplishments.[7]

It is also important to remember that, until Franklin Roosevelt's New Deal, it was not at all clear that market economies solved the productive investment CAP in a way that was favored by the main principle. The main principle evaluates changes by the extent to which they equitably promote life prospects. As Marx, among many others, pointed out, it was clear that laissez-faire capitalism promoted *some people's* life prospects, but the tendency of laissez-faire capitalism seemed to be to concentrate wealth among a small minority and to emiserate the great majority. The secret of the success of capitalism was that it motivated people to invest in providing goods and services that other people would *pay for*. The Achilles heel of capitalism was that it tended to promote life prospects in proportion to one's *ability to pay*. Thus, laissez-faire capitalism did not tend to promote the life prospects of those who had little to spend, of whom there were very many.

Marx thought that the solution was the abolition of private property, but the result of his solution was an economic system that took the exploitation of labor to a level unimagined in laissez-faire capitalism—for example, in Stalin's labor camps or on Mao's communal farms. To improve on laissez-faire capitalism, it was necessary not to abolish property and contract rights, but to redefine them. In order to understand why the process of redefinition has been largely endorsed by the main principle, we have to look at the rights themselves more deeply.

Market Systems as Self-Regulating Systems

I have said that economic rights are a solution to the productive investment CAP. In the next few sections I am going to explore the problem and the nature of the solution more fully.

For most of human history, the problem of productive investment was solved top-down, by institutions that dictated how people should act so as to make their actions productive. Sons learned the appropriate skills from their fathers and women from their mothers. Then in the past few thousand years, there developed centralized top-down command-and-control governance hierarchies in which decisions about production were made at the top by kings or emperors or other autocratic individuals or bodies and passed down the chain of command. This was, of course, the model adopted by the Marxist dictatorships.

Until the work of Adam Smith, there was practically no appreciation of the potential virtues of a decentralized model of economic decision making. Even in the nineteenth century, long after the work of Adam Smith, it was hard to believe that capitalism was a good system for productive investment, for there were far more business bankruptcies and other failures in capitalism

than in other economic systems. It was largely due to the work of Hayek (1960), who translated the Millian epistemology into economic terms, that the advantages of a decentralized economy came to be appreciated. I begin with Hayek's account and then add to it.

Hayek pointed out that markets allow everyone to benefit from the knowledge distributed over the entire community or society. To do this, the system must motivate people to share their knowledge in a form in which it will be useful to others. This it does by permitting people to produce goods or provide services for sale and permitting individuals to purchase those goods or services at mutually agreed upon prices. To explain the potential advantages of market economies, I am going to assume a very simplified, idealized model of such economies, the kind of model used in introductory economic textbooks. I use this model not to describe real-world market economies, but rather to articulate the goal by reference to which improvements in real-world market economies can be evaluated.

On the assumption, which I examine shortly, that voluntary exchanges improve the life prospects of both parties, markets establish a feedback mechanism that motivates people to attempt to promote other people's life prospects, because it rewards with monetary profits those who succeed in producing goods or services at a price that people are willing to pay and because it penalizes with monetary losses those who fail to do so. The model does not work because people are intrinsically motivated by money. It works because money represents the potential for employing goods and services produced by other people to promote one's own life prospects. Of course, there are other kinds of relationships with other people that promote life prospects: relations of family and friends. But market economies enable everyone to have their life prospects promoted by strangers whom they will never know. Think of all the thousands of people whose knowledge and creativity contributed to the production of the personal computer or the cell phone or the sewing machine at a reasonable price and of all the millions of people whose life prospects have been enhanced by the knowledge and creativity of those thousands of people they will never know.[8]

Although personal computers, cell phones, and sewing machines are all success stories, in a free market system, for every success there are multiple failures. In a market system, businesses that do a better job than their competitors of promoting the life prospects of their customers are rewarded and those that do a worse job are penalized, where the ultimate penalty is to go out of business.[9]

Because ideal market systems tend to encourage activities that promote life prospects and to discourage activities that reduce life prospects, they are a special kind of self-regulating system: a self-regulating system for promoting life prospects. They do not require a central authority to decide which enterprises should thrive and which should fail. No individual has the knowledge that would be necessary to make such determinations, but the system defined by laws of property, contracts, and civil liability makes them.[10]

A self-regulating system uses a feedback mechanism to improve itself without anyone having to direct the process. In a market system, the feedback mechanism is the system of voluntary exchanges, by which buyers' purchasing decisions reward, and thus encourage, the production of goods and services that promote their life prospects (with greater rewards going to those activities that best promote life prospects) and penalize, and thus discourage, and, over time, eliminate, the production of goods and services that do not promote life prospects (or that do so less well than available alternatives).[11] To see how this works, I begin by focusing on increases in life prospects. Equity comes later.

Market incentives play an important role in eliciting productive investment. Contrast a market system with an egalitarian one. In an egalitarian system, everyone would share equally in the productive activities. In such a system, everyone would get just as many resources to make movies as Stephen Spielberg. As a result, Stephen Spielberg would get to make only a few low budget movies. Because he is such a talented moviemaker, I imagine that even his low budget movies would still attract large audiences. This would not be true of most movies that would be made under the egalitarian system. People who are as unskilled at moviemaking as I am would be lucky if a few close family members and close friends were willing to sit through our movies.

In a market system, Spielberg's success in attracting audiences translates into profits and, as a result, he is able to attract more resources for making movies and he gets to make more movies with larger budgets than almost anyone else. This greatly promotes life prospects by comparison with the egalitarian system. Far more people enjoy many more movies and their life prospects are enhanced.

In a market system, Spielberg will be offered his *market value*, or, as Marx called it, his *commodity value*, for his services in making movies. One of the reasons that philosophers have been so antagonistic to market economies is that they find it morally offensive to evaluate people by their commodity value. They are right that commodity value is not any way to value a person. But a market system is not a system for determining objective value. It would be hard to find anything of more objective value to human beings than the air we breathe, but the commodity value of that air is zero.[12] A commodity value depends on scarcity as well as on contribution to life prospects.

In a market system, it is inevitable that there will be inequalities in income (i.e., commodity value), because of differences in what people are willing to pay for the products of one's activities. These differences are inevitable, not only because some businesses succeed and some fail, but also because not all activities make the same contribution to promoting other people's life prospects.

If an ideal market system is one that would reward individual activity in proportion to its effectiveness in promoting the life prospects of other people, then we can see why changes to a real-world system that led it to more closely

approximate this ideal would be approved by the main principle. In choosing activities with the highest commodity value, individuals would be promoting their own life prospects and also maximally promoting the life prospects of others.

Which others? Here we confront an issue that is fundamental to designing a market system that will be endorsed by the main principle. Ideal market systems are self-regulating systems for promoting life prospects in proportion to one's ability to pay. Because the feedback mechanism is the money generated by the purchasing decisions of buyers, a market system promotes the life prospects of buyers, in proportion to how much they buy. Of course, the main principle endorses practices on the basis of how well they *equitably* promote life prospects. So for a market system to earn the endorsement of the main principle, it must be so structured that promoting life prospects on the basis of ability to pay also equitably promotes life prospects. This is not an easy problem to solve. In this chapter, I consider many of the changes in U.S. law that have helped to solve it and suggest further improvements.

The Process by Which Property and Contract Law Have Improved in the United States

A libertarian system of property and contract law would have provisions for original acquisition of property and for voluntary transfers by psychologically competent adults, so long as there was no force or fraud. There has never been such a simple system of property and contract law, but nineteenth-century U.S. law provides a reasonable approximation. The historical story of how the nineteenth-century system evolved into the present-day system will enable me to show how the main principle applies to economic rights.

Posner characterizes the change as a change from a system of *caveat emptor* to *caveat venditor* (1983, 184). This is an apt description. Let us explore some of the elements of the change.[13]

Implied Warranties and Strict Liability

In sales contracts, the seller typically has more information about the items being sold than the buyer.[14] If the seller makes a false representation about the goods being sold, then the contract is void due to fraud. But what if a seller offers for sale garden hoses that he knows are water soluble. It would not occur to most buyers to include in the contract a provision that requires that the hoses not dissolve in water. Rather than allow the seller to benefit from this sort of information asymmetry, in the common law, there developed two kinds of *implied warranties*, an implied warranty of merchantability (e.g., that the items sold as garden hoses would function as garden hoses are reasonably expected to function) and an implied warranty of fitness

for a particular purpose (e.g., that ordinary garden hoses are not sold as high pressure hoses).

It seems unfair that sellers could benefit from these sorts of information asymmetries. However, the main principle does not evaluate individual transactions, but rather practices (cf. Crasswell 2001). In this case, the relevant practice would be the practice of imputing the implied warranties to contracts that did not explicitly contain them. To determine whether this would be endorsed by the main principle as a substantive practice, we have to consider the incentive effects of the two alternatives. It will help to have a general characterization of the kinds of cases involved. They are cases in which both parties consent *ex ante* to the sale, but only because there is an *information* asymmetry—that is, there is relevant information that one party has that the other lacks. If the ignorant party had had that information, it would not have consented to the sale. I refer to contracts that raise the life prospects of only one party as *one-sided contracts*.

Note how this example illustrates the subclass logic of moral and legal norms. Even if it is true that voluntary agreements generally increase the life prospects of both parties, a subclass of voluntary agreements, one-sided contracts, increase the life prospects of only one party. For a judge in a system of caveat emptor to uphold the status quo would be to uphold a system that provides incentives for parties to create and profit from information asymmetries and, thus, to increase the number of one-sided contracts. A judge who adopts the doctrines of implied warranty would reduce, if not eliminate, that incentive. So one effect of the change would be to reduce the number of one-sided contracts.

If we consider only the incentives for parties who might be tempted to profit from information asymmetries, we miss an important aspect of the comparison. In a *caveat emptor* system, information asymmetries increase the expected costs of contracts to potential buyers and thus make contracts less attractive. Reducing the probability of sellers' profiting from information asymmetries will decrease the expected costs for buyers, with the result that buyers will be more motivated to enter into contracts and thus there will be more of them. So the effect of the change to the status quo to add the two implied warranties to contract law would not merely be a *reduction* in the number of one-sided contracts, but what would almost surely be an even greater *increase* in the number of two-sided contracts, as potential buyers had less reason to fear being party to a one-sided contract.

What about the costs of the change? Presumably, there would be some increase in litigation. Notice that if the solution were to void contracts involving information asymmetries, we would expect there to be disputes about what the seller did or did not know. However, the implied warranty doctrines avoid this litigation expense by imposing on the seller strict liability (liability without fault) for the quality of the good sold.[15] From a nonconsequentialist perspective, it is hard to see how liability without fault could ever be justified. But from a consequentialist perspective, it is easy to see how

it could be. A standard of strict liability is an improvement under the main principle because of its incentive effects. It motivates the person who can do so at least cost, the seller, to make the determinations necessary to assure that the goods covered by the contract will perform to the expectations of the other party, the buyer.

The problem of information asymmetries played an important role in the development of another common law strict liability doctrine, the strict liability of manufacturers/sellers for defective products. This doctrine was announced by the Supreme Court of New Jersey in a 1960 case, *Henningsen v. Bloomfield Motors, Inc. and Chrysler Corporation*.[16] The Henningsens bought a new automobile from their local Chrysler dealer. The contract of sale limited the damages for defective parts to replacement of the defective part. There was a defect in the steering wheel. Ten days after the sale, while Mrs. Henningsen was driving the car, the steering wheel broke, she lost complete control of the vehicle and collided with a wall, totaling the car. Chrysler maintained that they were responsible only to replace the defective steering mechanism.

The court disagreed and held Chrysler liable for the full loss, on the basis of an implied warranty of merchantability. However, even at the time, the court was aware that it was enunciating a new doctrine of strict liability, because this new liability for defective products did not depend on privity of contract.[17]

What was most surprising was not that the court held the manufacturer liable for damages from a defective product, but that the liability did not depend on fault. What the court did was to make the manufacturer/seller an insurer not just of the buyer, but of the public at large, for harms caused by defective products. Thus, the new doctrine was not so much an extension of the implied warranties of contract law as it was an extension of the doctrine of strict liability for hazardous activities, which required no privity.[18]

Why would the court require the manufacturer/seller to insure the public against defective products? There are many possible answers to this question. For one thing, because the manufacturer is the person most knowledgeable about potential hazards and in the best position to prevent them, making the manufacturer an insurer provides it with an incentive to make its products safer. But there is a deeper answer that ties into my earlier discussion of markets as selection processes. Ideally, markets should reward goods and services that promote life prospects and penalize those that do not. When liability extends only to those with privity of contract, there is a potential for the market to reward products and services that promote the life prospects of buyer and seller, no matter how much they reduce the life prospects of third parties. Eliminating the privity requirement is a way of internalizing these external costs. Because product liability law applies only to products that cause severe harm, it is a way of internalizing the most severe costs. Of course, a full accounting would require also internalizing the external benefits. But, as a general rule of thumb, if the most severe costs are internalized, the other less severe noninternalized costs would be expected to be offset by the other noninternalized benefits. So the law of product liability can be seen

as an attempt to realize the full promise of a self-regulating, decentralized market system that would reward activities that increase life prospects overall and eliminate activities that reduce them, all without the need for a central authority to decide which activities fall in each category.[19]

Each of these examples of strict liability involves a limitation on voluntary contracts. These limitations have been criticized as paternalistic—that is, as the government overruling people's own judgments about what is good for them. It is important to see that none of these limitations depends on a paternalistic justification.

Consider first the implied warranties in contract law. These implied warranties protect the buyer from one-sided contracts. No reasonable buyer wants to be a party to a one-sided contract, so the law is not paternalistic toward the buyers. It gives effect to the buyers' judgments about what is good for them; it does not overrule it. Similarly, strict products liability is not paternalistic, because even if both the buyer and seller would prefer not to have insurance against severe harms caused by defective products, the protection of third parties is a nonpaternalistic justification. There is another reason that systems of mandatory insurance against severe harm are not paternalistic, even when those who are potentially harmed would prefer not to have the insurance. I discuss this issue more fully when I discuss social insurance rights in chapter 11.[20]

Mandatory Disclosure

Implied warranties are a judicial solution to the problem of information asymmetries. A common legislative solution is to enact mandatory disclosure laws. Mandatory disclosure laws are an important part of consumer protection. These cover the requirements of informed consent that must be obtained before medical procedures, the requirement that manufacturers list the ingredients in food products on the label, the requirement that sellers complete disclosure forms and reveal hidden defects to potential buyers in a home sale, the requirement that car mechanics provide a written estimate before doing repairs, the requirement that investors be provided with a prospectus before investing, and so forth. From the point of view of the main principle, these requirements can be seen to reduce the probability of one-sided agreements. What would a nonconsequentialist say about such requirements? I take up that question shortly, after I have completed my brief overview of the evolution of the law of voluntary agreements.

Win-Win Agreements and the Replacement of Ex Ante *with* Ex Post *Consent*

Consumer contracts are a subclass of agreements that I call *win-win agreements*, because such agreements typically increase both parties' life prospects not only *ex ante* (at the time of the initial agreement), but also

ex post after the transfers of goods or services have taken place and each party is in a position to better evaluate the consequences of the agreement. This suggests that the law might be modified to require not only mutual consent *ex ante*, but also *ex post*. This would be an effective way of preventing one-sided agreements. Somewhat surprisingly, the law has made significant moves in this direction. In many jurisdictions, the law now gives buyers the option of voiding a house sale within some specified number of days after signing the contract. The Medicare rules have been changed so that Medicare no longer pays for medical care that is unsuccessful.

What is even more surprising is that the replacement of *ex ante* with *ex post* consent in win-win contracts has gone far beyond what is mandated by law. The mere existence of Better Business Bureau reports on customer complaints has provided most businesses with a powerful incentive to address the concerns of dissatisfied customers. In addition, most retailers accept returns within a specified amount of time for any reason at all.[21] Some even offer purchase insurance, promising to pay the difference (or, in some cases, more than the difference) if the buyer finds the same item advertised at a lower price within a specified time period. Because of the potential for abuse of such policies by customers, it would have been impossible to know in advance if such policies would be profitable. Now that we know they are profitable, the question is, why?

The answer is the very much the same as the answer to the question of why the main principle endorses the implied warranties of contract law: first, such policies almost eliminate any incentive for sellers to propose one-sided sales contracts; second, eliminating the risk of one-sided sales contracts has the effect of motivating many more sales, because buyers no longer must assume the risk of a one-sided sale. Thus, such policies increase the number of two-sided sales. Of course, this assumes that the increased business is enough to offset the losses to sellers from returns, especially by those who abuse the system.

Thus, in brief, we can see how the United States has evolved from a system of *caveat emptor* to a system of *caveat venditor*, at least for win-win contracts. I discuss win-lose contracts in a footnote.[22]

The Significance of Consent

In the preceding section, I only skimmed the surface of changes in the law of contracts from the nineteenth century to the present. I have said nothing about the many other judicial exceptions to voluntary contracts, including exceptions for impracticability, frustration, mutual and unilateral mistake, illegality, excessively high liquidated damages, and, as discussed previously, unconscionability.[23] Under unconscionability, I include slavery contracts and contracts of indentured servitude. I discuss the unconscionability exception in the law under the category of legal solutions to CAPs shortly. I have

also said nothing about legal limits on property rights that go beyond the mere requirements of preventing harm to others, including prohibitions on monopolies and other anticompetitive practices, the doctrine of adverse possession, zoning restrictions, and limitations on perpetuities. All of these examples are easily seen to be improvements when evaluated under the main principle. For the sake of brevity, I focus my discussion here on the three changes discussed in the previous section.

In the previous section, I discussed how the doctrines of strict liability, including the implied warranties of contract law and products liability, mandatory disclosure requirements, and substituting *ex post* for *ex ante* consent in win-win contracts, would be endorsed by the main principle. On what grounds would a nonconsequentialist endorse them? I limit my consideration to nonconsequentialist accounts based on the value of autonomy, or some reasonable facsimile thereof. Of course, it is always open to explain them by reference to some sort of hypothetical consent. I have already addressed such views in chapter 2. If the reason that they would be hypothetically consented to is that they promote life prospects, then the hypothetical consent test is just a proxy for consequentialism.

On the consequentialist account, each of these examples illustrates how, by changing the legal framework, it is possible to enhance the tendency of voluntary agreements to promote life prospects. On the consequentialist account, because voluntary agreements are themselves favored because of their tendency to promote life prospects, constraints on voluntary agreements that enhance their tendency to promote life prospects are improvements. The challenge to the nonconsequentialist champion of autonomy is to provide a nonconsequentialist account of the value of autonomy that can explain why these kinds of *constraints* on voluntary agreements promote the value of autonomy.

Of course, it is open to the nonconsequentialist to oppose these constraints on voluntary agreements on the grounds that they reduce autonomy. This is the position of a libertarian who advocates a system of *caveat emptor*. To the libertarian, it is a moral problem that people aren't free to induce others to enter into one-sided contracts or that manufacturers aren't free not to compensate those injured by their defective products, if they are not at fault for the defects,[24] or that sellers of homes would be required to obtain consent not only *ex ante* consent, but also *ex post*. As discussed in chapter 2, the libertarian would also think it was a problem that people aren't free to enter into unconscionable contracts, including contracts of slavery and indentured servitude. There is really not much more I can say to a libertarian.

The nonconsequentialist I wish to address is one who agrees that all three of the changes discussed in the previous section are moral improvements, but that my consequentialist explanation of why they are improvements is mistaken: someone who holds that the true explanation is that they promote autonomy, not life prospects or some other measure of well-being.

How exactly do they promote autonomy, understood nonconsequentially? On the Kantian *metaphysical conception* of autonomy, there is no answer to this question. *Metaphysical autonomy* requires acting without being caused. How much information we have before making a choice is irrelevant to its metaphysical status. Scanlon proposes a nonmetaphysical but still nonconsequentialist account of autonomy, according to which the value of choice (i.e., autonomy) is understood in terms of the significance of having outcomes depend on one's choices (1998).

This is an initially puzzling suggestion. One would think that a nonconsequentialist would agree with the platitude "It's the thought that counts." But for Scanlon, good intentions are not enough. Outcomes matter. So why doesn't Scanlon identify himself as some kind of consequentialist? Scanlon thinks that outcomes matter in a way that cannot be explained in a consequentialist framework such as mine. Scanlon thinks that choices have two kinds of value, instrumental and noninstrumental.[25] *Instrumental* value is the most familiar (1998, 251–254). If outcomes depend on our choices, our choices can enable us to bring about valuable outcomes. To the extent that the valuable outcomes are understood in terms of (appropriately distributed) well-being, a consequentialist account can easily account for this element in the value of choice.

The *noninstrumental* value of choice can be illustrated by gifts. Under reasonable assumptions, economic theory implies that for a given amount of money, we could not buy a gift that would be better than a gift of that amount of cash (Waldfogel, 1993). But gifts can have value that goes beyond their cash value, because they can be an expression of one's thoughtfulness or devotion or commitment.

If Scanlon were correct that the consequentialist could explain only the instrumental value of choice, then he would be right that any consequentialist account of the significance of choice would have to be incomplete. He might even allow that economic relations were typically instrumental relations and thus that a consequentialist account of the appropriate economic relations might be possible. But he would insist the value of many kinds of human activities and relationships could not be understood on the basis of instrumental value alone and thus that consequentialism could not explain the value of choice in such activities and relationships.

I believe that Scanlon's argument is effective against ground-level or direct consequentialism. But what about my meta-level consequentialism? Can my meta-level consequentialism explain the noninstrumental value of choice? Consider the example of the gift. The noninstrumental value of gifts can play an important role in relationships by communicating such things as commitment, loyalty, esteem, devotion, and concern. All of these attitudes and more can be expressed by a personally crafted gift, though none of them would typically be expressed by substituting the monetary value of the gift.

But we should be puzzled by Scanlon's assumption that such values cannot be explained in terms of well-being. After all, relationships based on

commitment, loyalty, esteem, devotion, and concern do clearly make an important contribution to well-being. Think how bleak life would be without such relationships.

To see that their contribution to well-being could be the basis of their noninstrumental value, imagine a world in which adults who entered into multiple short-term superficial relationships were happier than those who entered into a smaller number of long-term committed relationships and that adults in the former group produced children who were happier than the children of adults in the latter group. Is there some value of committed relationships that would make it reasonable for people to continue to value them above uncommitted superficial relationships, when they were no longer a source of well-being?[26]

This suggests to me that it is because the noninstrumental value of choice indirectly contributes to well-being that it qualifies as a value at all. So I conclude that Scanlon's account of the noninstrumental value of choice is compatible with an indirect consequentialist explanation of its noninstrumental value.

What's So Good about Voluntary Consent? Smith versus Kant

Kant is the prime representative of the tradition that explains the significance of voluntary consent on the basis of the value of autonomy. Adam Smith is probably the prime representative of the consequentialist tradition that explains the significance of voluntary consent on the basis of its role in promoting well-being.

Consequentialists typically assume that voluntary agreements tend to promote the well-being of both parties. But why? And why would the kinds of changes that have transformed the U.S. system from a system of *caveat emptor* to *caveat venditor* tend to better promote well-being? The answer is that, with some qualifications to be mentioned shortly, the judgments of adults about what is good for them are generally reliable and typically more reliable than the judgments of others. This is the *claim of first-person authority*. It seemed obviously false to most educated people when Mill first made it ([1859], 86–87), and it is still controversial today. In the first volume I considered reasons for thinking that it is true (2005, 123–128). One further reason that I did not mention there is that markets would fail if it were not at least approximately true. In fact, it is only a slight exaggeration to say that this foundation of human rights, the claim of first-person authority, was discovered by accident, in large part due to the unplanned and unexpected success of markets.

In the first volume (Talbott 2005, chap. 6), I explained the role of *autonomy rights* as enabling people to develop the capacity for making reliable judgments of what is good for them (i.e., good judgment) and to exercise it.

Idealized markets motivate sellers to produce goods and services that will promote the life prospects of prospective buyers. Real-world markets create the incentive for sellers to get buyers to believe that their goods or services will promote their life prospects, *regardless of whether that belief is true*. Thus, real-world markets do not necessarily reward sellers who produce goods and services that truly are good for those who purchase them. They may actually reward sellers who are good at producing false beliefs—especially false beliefs of buyers about what is good for them. These sellers induce buyers into one-sided contracts, for example, by taking advantage of information asymmetries.

Because markets reward sellers who can induce buyers into bad bargains, successful markets depend on there being a framework within which buyers' judgments are generally reliable. It is possible to establish such a framework without government involvement, if there are ongoing relationships (Greif, 2006). A seller would not usually be motivated to take advantage of a buyer in a single transaction if that would jeopardize a large enough number of potential future transactions. But it typically requires a government to provide a framework for sales between parties who have no ongoing relationship. So in an economic context, the claim of first-person authority must be made relative to a background framework. The claim is that if an appropriate background framework is established, including the basic human rights, then people's judgments about what is good for them will tend to be reliable. If this were not true, market economies would collapse.

It is not only buyers who have an interest in establishing a framework in which buyers' judgments about their own good are reliable. Reputable sellers do also. As Marx [1867] reports, reputable London bakers agitated for government regulations governing the contents of bread, because otherwise they would not be able to compete with bakers who replaced some of the flour with sand. In the United States in 2009, an outbreak of salmonella contamination in peanut butter from one manufacturer reduced overall peanut butter consumption by 25%. The reputable peanut butter companies would have benefited from more strict FDA oversight.[27]

The three kinds of changes to contract law discussed above—the implied warranties and strict liability for product defects, mandatory disclosure laws, and the substitution of *ex post* for *ex ante* consent for win-win contracts—are all changes that reduce the incidence of one-sided agreements by making the buyer's judgments more reliable. It is obvious how the implied warranties and products liability law reduce the incidence of one-sided agreements. It is also obvious that mandatory disclosure laws reduce the incidence of one-sided agreements, but it is useful to say something about how they do so.

Mandatory disclosure laws require disclosure of information by sellers not because more information is always better than less. Too much information can generate information overload and interfere with good judgment.

Mandatory disclosure laws target relevant information. Relevant to what? Relevant to the reliability of the decision at issue.

How does information improve the reliability of judgment? Here we are reminded again of the reference class logic of reasoning. I expect a jar of peanut butter to promote my life prospects, but not a jar of peanut butter containing salmonella. Conditional probabilities can change with the addition of more information. Relevant information is simply information that has the potential to alter the relevant conditional probabilities.

What about the substitution of *ex post* for *ex ante* consent in win-win contracts? Here again, it seems obvious that one's judgment about the wisdom of buying a house would be more reliable after spending some time in the house. This general phenomenon, that *ex post* judgments about one's own good are more generally reliable than *ex ante* judgments, plays an important role in my account of rights against paternalism in chapter 12.

Almost all of the changes discussed so far can be understood as changes to *caveat emptor* that reduce the probability of one-sided contracts by providing a framework in which buyers' judgments are more reliable. That is not the only kind of change that might be endorsed by the main principle. For example, in some jurisdictions, a party to a contract may choose to pay compensation rather than to perform on the contract. Such *efficient breach* conflicts with the ground-level moral norm that we should keep our promises. Posner (2007) is the most well known advocate of efficient breach. Fried (1981) argues against efficient breach on the basis of the ground-level moral norm of promise keeping. But if the doctrine of efficient breach promoted everyone's life prospects, it would easily pass any hypothetical consent test and be endorsed by the main principle. Whether it would promote life prospects turns out to be difficult to determine, because of the great variety of incentive effects (Craswell 2001, 26–32).

Common Law Adjudication and the Main Principle

In this chapter, I have considered a number of changes in the law, including judge-made law in a common law tradition, that are endorsed by the main principle. It is important to emphasize that I am not claiming that judges in a common law tradition should apply the main principle in their adjudications. This would be to make the mistake of taking the main principle as a ground-level principle of adjudication. As I explained in chapter 5, judges do a better job of satisfying the main principle if they don't directly apply it. But the fact that the principles that judges do apply all have exceptions is an indication that their reasoning has the reference class logic that it would have if it were explained by the main principle.

I discuss the role of the main principle in legislators' deliberations in chapter 10. Now I turn my attention to other kinds of economic rights.

Prohibitions on Slavery and Indentured Servitude and Other Minimum Wage Legislation

It is very difficult to provide a nonconsequentialist explanation of prohibitions on slavery. If the goal of autonomy rights is to promote autonomy, why shouldn't we be permitted to autonomously surrender some or all of our rights? It is this logic that drives Nozick (1974, 331), G. Dworkin (1983, 111), and Thomson (1990, 283) to the view that people should be free to choose to be slaves. Mill's argument for autonomy rights in *On Liberty* seemed to be taking him in that direction, but he balked and upheld a prohibition on slavery contracts ([1859], 115–116). But it is clear from Mill's argument that he failed to appreciate why slavery contracts should be prohibited. His argument was a general argument against all contractual limitations on liberty. This could not be correct. One of the primary ways that we promote our life prospects is by contractually limiting our liberty. So if there is something problematic about slavery contracts, it cannot be simply that they are contractual limitations on liberty.

There are two potential problems with slavery contracts. I discuss the first here and the second in chapter 13. The first problem with slavery contracts is that the decision to enter into them is typically a CAP for the potential slaves. It would be hard to find anyone who aspires to a career as a slave. So, as illustrated by the example of the medical researcher Marie, in chapter 2, prohibiting slavery contracts has the potential to benefit *all* the potential slaves. All of them would be better off if they were "forced" to work for wages rather than "allowed" to enter into slavery contracts. This is the logic of a CAP. In a CAP, the problem is not that the individuals' judgments are unreliable, but rather that each individual's acting so as to promote her own life prospects will, collectively, reduce the life prospects of all members of the relevant group. It was this very logic that Marx thought was integral to capitalism. If workers were faced with starvation, it would be rational for them to choose wage slavery to the capitalist. It is important to remember that the brand of laissez-faire capitalism that Marx opposed did not include any minimum wage or maximum hours laws, no worker health and safety laws, no labor unions or collective bargaining, not even child labor laws. All of these laws can be understood as solutions to workers' CAPs.[28]

They are *workers'* CAPs, because solving them benefits the workers, not the capitalists. It is easy to see that the main principle would endorse prohibitions on slavery and indentured servitude, even if they don't improve the life prospects of potential masters. Because the main principle favors equity, it gives priority to the life prospects of the less well off, in this case, the potential slaves. Although the case is so clear-cut that no test is necessary, the expanded original position (EOP) reinforces this result. From behind the veil of ignorance in the EOP, no one would be willing to accept that some would have their life prospects reduced to the level of slaves in order that the life prospects of others could be raised to the level of slave masters.

Minimum wage legislation is more controversial, but in theory the logic is the same. Many economists oppose all minimum wage legislation. One reason is that such laws are inefficient. Their effect is to increase prices above what their equilibrium level would be. But efficiency is not the main principle's criterion of improvement. The goal should be improving life prospects, with extra weight to the life prospects of the least well off. If the goal were efficiency, then it might be necessary to reevaluate prohibitions on slavery, because systems of slavery can be efficient (e.g., Fogel and Engerman 1974; Satz 2009). Another reason that many economists oppose minimum wage laws is that such laws can be expected to increase unemployment. If our concern is with the life prospects of the least well off, shouldn't we give priority to the life prospects of those who would lose their jobs if the minimum wage were increased?[29]

It is true that we should give priority to those who are laid off, but doing so does not necessarily favor the opponents of minimum wage legislation. Even if minimum wage legislation causes some low-wage workers to lose their jobs, it need not reduce their *life* prospects. For example, it is quite plausible that legislation increasing the minimum wage might increase the lifetime earnings of all minimum wage workers, even if some of them lost their jobs as an immediate consequence of the legislation. After all, they would have the rest of their lives to make up the loss.

Another argument against minimum wage legislation is that there is a danger that the minimum will be set too high. This is a genuine concern that needs to be carefully addressed. However, it is hardly a reason to oppose all minimum wage legislation, because it has to be balanced by the danger that the minimum will be set too low.

One final argument against a minimum wage is that there are better alternatives for raising the earnings of low-wage workers. In order to evaluate this argument, it is useful to consider a promising alternative.

The Negative Income Tax

The example of minimum wage legislation is a reminder that there is a gap between the standard defense of a market system as efficient and the requirements of the main principle. The main principle requires that social practices equitably promote life prospects. Market economies promote life prospects, but there is no guarantee that they will do so equitably. Indeed, a pure market system promotes life prospects in accordance with people's *ability to pay*. As a general rule, those with more to spend have higher life prospects. The greatest challenge to designing a market system that would be endorsed by the main principle is to design a system that effectively motivates individuals to choose to engage in productive activity while equitably distributing the benefits of productive activity measured by life prospects. Because in a pure market system, people's incomes are equal to their commodity value, there would

be no problem if people's commodity values were roughly equal. Then everyone would be able to purchase roughly equal shares of the social product.

But in any real-world market system, there will be great inequalities in people's commodity value. In 2008, Steven Spielberg's commodity value was $330 million; the median commodity value of a schoolteacher was around $40,000; and the commodity value of a full-time minimum wage worker was approximately $13,000. These disparities in commodity value translate into a very large difference in life prospects. It is true that the main principle could endorse such great disparities, if, for example, the trickle-down defense of capitalism were true and it really was the case that allowing such inequalities benefited everyone, and that there was no alternative system that would increase the life prospects of those near the bottom. But this is very implausible.

Is there an alternative to a system that pays everyone their commodity value that does a better job of equitably promoting life prospects? We have already discussed why Marx's egalitarian socialist alternative fails. In a market economy, everyone is encouraged to use their talents and creativity in ways that promote other people's life prospects. In an egalitarian socialist system, the government decides who produces what and talent and creativity languish.

Are there any alternatives between these two extremes? As it happens, there are lots of alternatives. The most promising alternative combines a progressive income tax on high earners with a negative income tax (e.g., an earned income credit) for low earners.[30] In the United States in 2009, the highest marginal income tax rate is 28%. A minimum wage earner with two children qualified for the maximum earned income credit of $4,716. If Steven Spielberg's income were taxed at the highest marginal tax rate, he would pay taxes of $92 million, which would fund a maximum earned income credit for nearly 20,000 low income workers.

I think it is easy to show that the main principle would endorse a transfer of this kind. In fact, I believe that the main principle would endorse a marginal tax rate of at least 50% and a negative income tax that would at least double the income of minimum wage earners. How could this be? Wouldn't transfers of this size make the system as a whole inefficient? I believe there is a great potential to increase equity with little effect on incentives for productive investment. To see why, it is necessary to say something more about the role of commodity values in a market system.

In a market system, commodity values are given by market prices. For human commodities, these prices are salaries or annual pay. In an ideal market system, salaries function as signals. The higher the salary, the more productive the job. So long as human commodities choose the job with the highest salary, the result will be efficient, regardless of whether they actually receive as income the full salary for their services. Because a graduated income tax does not change the ranking of alternative jobs or careers, even with a graduated income tax of, say, 50%, salaries would perform the same signaling function. If the salaries of jobs order them in terms of productivity, cutting all the salaries by half will preserve that order. If Steven Spielberg has a choice between two movie projects, one that pays $10 million and another that pays

$5 million, an income tax rate of 50% would reduce his actual income to $5 million or $2.5 million, but the order would be preserved. So a graduated income tax does not interfere with the signaling function of prices.[31]

Another advantage of a graduated income tax on high-wage earners is that it maintains the psychic rewards that they derive from knowing their commodity value. When Steven Spielberg and Oprah Winfrey get together, they can compare their commodity value if they wish ($330 million to $225 million). Because human motivation is so largely comparative, if Spielberg's motivation is to outdo Winfrey (and hers is to outdo him), they will both choose to maximize their commodity value, regardless of whether they receive 100% of their commodity value or only 50% of it in income. (In the former case, he outranks her $330 million to $225 million; in the latter, he outranks her $165 million to $112.5 million.) Thus, to the extent that human motivation is competitive, the 50% tax rate will have no effect on the motivation of high-wage earners to engage in their most productive activity.

Of course, human motivation is not entirely competitive, so any increase in the tax rate will potentially have some effects on motivation. And yet we have very little information about what those effects are. Undoubtedly, there are tax rates on high-wage earners that are low enough, say 10%, that their effects on motivation would undoubtedly be negligible. And there are tax rates high enough, say 99.9%, that they would have seriously negative effects on motivation. What about tax rates in between? All we have is anecdotal evidence. But that anecdotal evidence does not support the common assumption that increasing the current maximum marginal tax rate of 28% would reduce the productivity of high-wage earners. For example, we know that the highest marginal income tax rates in the United States were above 90% in the 1950s and that was a period of high economic growth. In 2009, the highest marginal tax rate in China was 45%, and China had one of the world's fastest growing economies. It is this kind of evidence that makes me think that a 50% marginal tax rate on high earners is a conservative estimate of where to set the upper marginal tax rate so as not to have a significant adverse effect on productivity.[32] And I see no reason to think that using the increased tax receipts to fund a larger negative income tax for low-wage earners would negatively affect productivity either. No one could plausibly hold that doubling the income of movie directors from $165 million to $330 would enhance productivity but that doubling the income of minimum wage earners would not.

Objections to Redistributive Taxation

The Forced Labor Objection

Nozick raised an influential objection to this kind of redistribution. He argued it was morally equivalent to forced labor (1974, 169–170). To make his point, he asked why someone, call him *Donald*, who needs money to get enjoyment

(e.g., who enjoys going to the movies) should be taxed on the earnings that he makes to be able to pay for enjoyment, whereas someone, call him *Jerry*, who does not need money to get enjoyment (e.g., who enjoys watching a sunset) need not have any earnings that would be taxed. He pointed out that the tax will make it necessary for Donald to work longer to get money for the movies and suggested that equity would require forcing Jerry to do some extra work, also.

But Nozick misunderstood the equity problem that the negative income tax is designed to solve. No policy could assure everyone the same level of happiness. The equity problem that the negative income tax is designed to solve is the inequity in life prospects generated by differences in annual earnings due to differences in people's commodity value, regardless of whether they enjoy sunsets or going to the movies. The proper comparison is not between Donald and Jerry, but between Donald and Archie. Both Donald and Archie enjoy movies, so they both need money to do what they enjoy. Because people are willing to pay more for Donald's services than for Archie's, Donald has a higher commodity value than Archie. In one hour Donald can make enough money to go to 100 movies, whereas Archie can make enough to go to only one movie. Other things being equal, the main principle would endorse a social practice of redistribution that reduced Donald's income to the equivalent of 50 movies per hour and increased the income of Archie and 49 other low-wage workers to the equivalent of 2 movies per hour, if the transfer did not alter their motivation to choose a job that maximized their commodity value.

Only if Nozick could make a serious argument that people with lower commodity value (and thus, lower earnings) generally have better life prospects than people with higher commodity value (and higher earnings), would he have a basis for arguing that it would be inequitable to reduce the very large inequalities that currently exist in earnings. Of course, Nozick would never have made such a claim. But if a negative income tax is aimed at making life prospects more equitable, then the analogy to forced labor is mistaken. Such a tax is *not* like forced labor. It is simply a way of making the monetary return to labor more equitable.

The Lifetime Earnings Objection

Another objection to the negative income tax proposal is that it is based on a mistaken analysis. If redistribution is based on annual income, it will not take account of the fact that younger people tend to have lower salaries early in their careers and higher salaries later. Viewed over the course of a complete life, such differences in income raise no problem of equity.

The problem with this argument is that when comparisons are made in terms of lifetime rather than annual earnings, there are still serious inequities.[33] The negative income tax proposal would undoubtedly benefit some people who, over the course of their lifetimes, would do well without it. But it will benefit even more those whose lifetime income is significantly below the median and average.

The Hard Work Objection

Another moral objection to this sort of redistribution is that Steven Spielberg, Oprah Winfrey and others with high commodity value have a moral right to receive their commodity value. For example, I am sure that Spielberg and Winfrey work very hard. They may even have started out working at low wages before becoming successes. Because they were willing to work hard, the argument goes, they deserve the rewards of their hard work.

This can't be right. It implies that the schoolteacher who worked long, hard hours for years at low wages would also deserve to be paid millions of dollars, just like Spielberg or Winfrey. The only plausible avenue for establishing Spielberg and Winfrey's entitlement to their income is to take into consideration the value to others of what they produce, because there are far too many people who work as hard and long as they do with much more meager compensation.

If there are lots of other people who work just as long and hard as Spielberg and Winfrey, this suggests that it is a mistake to say that Spielberg and Winfrey morally *deserve* such larger rewards. But Nozick does not think this settles the matter. In his terms, even if it is true that Steven Spielberg does not morally *deserve* to have a commodity value almost ten thousand times that of a schoolteacher, he is morally *entitled* to receive his commodity value (1974, 224–227). The puzzle is to explain why he is entitled to it. Nozick's answer is a historical one. When people acquire property in morally appropriate ways and transfer it by voluntary agreements, they are entitled to what they end up with, however great the inequalities are. This is an initially attractive moral view. But we know there is something wrong with it when we see that it implies that Marie, the medical researcher, could be morally entitled to enslave all of humanity.

These potential counterexamples point to a deeper problem with Nozick's idea of entitlement: Commodity value is a descriptive concept. It is determined by facts such as people's preferences, scarcity of talents, and opportunities to employ one's talents. But entitlement is a moral notion. How could the purely descriptive fact that Steven Spielberg has a commodity value of $330 million produce a moral entitlement to that income? The most common answer to that question is that the distribution resulted from voluntary choices. This position is much more plausible if there is some requirement that everyone start from a fair starting point.

Fair Starting Point Theories of Justice

Dworkin (2000) distinguishes between *starting-gate theories* of justice, which define a fair starting point in terms of resources and then endorse whatever distribution is generated by voluntary transactions, from his own view, which allows the resources in the starting point to include insurance against future

contingencies—disability, unemployment, or what he calls *underemployment*, which might usefully be thought of as insurance against low earning power. Because Dworkin evaluates the equality of resources *ex ante*, at a single time slice, it qualifies as a *starting point theory* in my sense. Any theory on which justice depends only on a single-time-slice distribution of anything is a fair starting point theory in my sense, unless the single time slice is at the end of life.[34]

Fair starting point theories attempt to capture an important moral intuition: that it is not fair that some people start life with great advantages while others begin with great disadvantages. This is a genuine concern of equity, and I address it in chapter 11. What is unsatisfactory about the fair starting point theories is that they hold that this is the *only* problem of equity. The implication is that if the starting point is just, then because voluntary transactions preserve justice, the results will also be just, however unequal they might be.

Dworkin's theory softens this seemingly harsh result by including insurance in his fair starting point. In addition to insurance against very bad outcomes (starvation, disability, etc.), Dworkin allows for what might be called *low-wage insurance*, that would provide additional income to low-wage earners, perhaps in the form of a negative income tax (2000, 97). Thus, Dworkin's proposal has the effect of guaranteeing a baseline of earnings to everyone. This is a welcome addition to traditional fair starting point theories. However, it has the consequence that if everyone is above the baseline level, there are no further requirements of equity. Thus, Dworkin reaches a qualified agreement with Nozick on the Wilt Chamberlain example (which I discussed in chapter 3). On Dworkin's account, if everyone has insurance that keeps them above the baseline level, there are no problems of equity, no matter how great the difference between the earnings of Wilt and the others (2000, 111).

This kind of case requires us to adjudicate between the two conceptions of the significance of choice discussed above. Do voluntary choices have the inherent power to transfer entitlements? Or do they owe that power to their being a part of a system that equitably promotes life prospects? It is hard to see why the inherent value of choice per se could rule out a redistributive system that operated by way of a graduated income tax. Suppose that such a redistributive system were in place. The ticket buyers would still *choose* to pay to see Wilt play. (There actually might be more of them, because the graduated income tax on high-wage earners would have been used to increase the earnings of lower wage earners, even if they were above the baseline level.) Wilt would still *choose* to play. The choices would all be the same. Only the distribution of income would be different. Why would justice require that Wilt have so much more than everyone else, when there was an alternative social practice of progressive taxation in which the choices would all be the same but the distribution of well-being would be more equitable?

The Illusion of Entitlement

It is important to realize that simply treating people as though they are entitled to their commodity value creates the powerful illusion that they are entitled to it. This is the illusion that Murphy and Nagel called the "myth" of ownership (2002). Those of us in the United States live in a country in which that illusion exerts a powerful influence in public policy discussions. But the power of the illusion is easily broken. To break it, all that is necessary is to go behind the veil of ignorance in the EOP and consider whether there would be agreement that people were entitled to receive their entire commodity value, if there were an alternative that was effective in motivating productive activity, in which high-wage earners received only half their commodity value and low-wage earners received double theirs. Behind the veil of ignorance, high- and low-wage earners alike would choose the latter alternative to the former, because, in general, dollars will be more valuable to someone who has relatively few of them than to someone who has lots of them and, perhaps as important, income is one of the social bases of self-respect. Great inequalities of income undermine self-respect. This result would follow even if the parties in the EOP were assumed to be rationally self-interested, but that is not the basis on which I arrive at it. It is not the fact that we would choose the redistributive alternative if we were rationally self-interested and behind the veil of ignorance in the EOP that makes the result equitable. The EOP test is just a way of removing our biases so that we can recognize the inequity. To reduce the inequity, the redistributive system would be endorsed by the main principle.

Negative Income Tax Plus Minimum Wage

Would a negative income tax be superior to minimum wage legislation as a way to promote equity? I suspect that some combination of the two would be better than either in isolation. The reason is simple. If there is no minimum wage, increases in the negative income tax could easily be offset by corresponding reductions in wages. Of course, employers could not reduce wages below zero, but it would be unfortunate if the effects of increasing the negative income tax were diluted by reductions in wages. To avoid such an outcome, I believe that it would be necessary to have both a negative income tax and a statutory minimum wage.

Other CAPs

I have suggested that the main principle would endorse solutions to a variety of workers' CAPs, including minimum wage and maximum hours legislation and occupational safety and health legislation. Labor unions are the classic

solution to workers' CAPs. The main principle would endorse at least a qualified right to unionize.[35]

Contracts of slavery and indentured servitude are examples of the broader class of unconscionable contracts. Legal theorists have struggled to formulate a theory to explain this category, because they have tried to do so by looking at the intentions of the parties or their rationality or their bargaining power or their duress (Benson 2001b, 185–195).[36] The difficulty for legal analysis is that, in a typical case, even though the price paid is excessive, it is generally not irrational for the victim to be willing to pay it. Any doctrine that looks only at the situation of the parties will not be an adequate doctrine of unconscionability. Here is why. Prohibiting unconscionable contracts must be evaluated as a practice, not simply on a case-by-case basis. If courts won't enforce unconscionable contracts, parties won't offer them. So the court deciding a contract case that raises the issue of unconscionability must consider not just fairness to the parties, but also the effects on future would-be parties if such contracts are not offered.

Consider two different cases. First, the case of Marie the medical researcher. If contracts of slavery are not enforced, Marie will charge a lower price for her cure. This is a solution to a CAP for the consumers, so the courts should hold that slavery contracts are unconscionable.

Now consider an example discussed by Hobbes [1651]. A nobleman is taken prisoner in war and is threatened with death. To avoid death, he promises to pay a large ransom when he returns home. After he returns home, should the contract be enforced? There could hardly be a contract extracted under greater duress. However, as Hobbes argues, if the captor had the right to put him to death, then he should be held to the contract. Why? Because if such contracts are not enforceable, in the future they will not be offered. But, unlike the case of Marie the medical researcher, in this case their not being offered won't make captives better off; it will make them worse off. They will be killed. Of course, this analysis assumes that the captor has the right to kill his captive. If so, Hobbes's analysis goes through. It would be endorsed by the main principle.

This is a reminder that the main principle does not apply on a case-by-case basis. It applies to practices, such as the practice of voiding consumer contracts as unconscionable. When applied to such a practice, it considers the incentive effects of a legal doctrine as a crucial part of the evaluation of it.

The main principle typically endorses solving consumer and worker CAPs. It does not typically endorse solving capitalists' CAPs. Price fixing and production quota agreements and other anticompetitive practices are solutions to capitalists' CAPs. These "solutions" enhance the life prospects of a relatively small number of the more well off by reducing the life prospects of a relatively larger number of the less well off, so they would never be endorsed by the main principle.[37]

Unemployment Insurance and Bankruptcy Protection

It may seem obvious that in a market economy, the main principle would endorse some sort system of unemployment insurance. But mandatory systems of social insurance such as unemployment insurance raise a puzzle. Why shouldn't unemployment insurance be voluntary? Workers could be given the option to contribute or not, depending on whether or not they thought the insurance was a good investment for them.

One potential response to this question would be to defend making insurance mandatory on paternalistic grounds—that is, by insisting that unemployment insurance would promote everyone's life prospects regardless of whether they agreed that it would. Then the main principle would endorse overruling the judgment of those who would decide not to buy the insurance and forcing them to buy it.

This is not a response that I can give to the question of how the main principle could endorse mandatory unemployment insurance, because I believe that, given the appropriate background conditions specified by the list of basic human rights, the claim of first-person authority is true and, thus, people are reliable judges of what is good for them. This is the basis for a right against paternalism, as I discuss in chapter 13. So if I am to maintain that the main principle would endorse a mandatory scheme of unemployment insurance, I need a nonpaternalistic rationale for that result. I provide one for a variety of systems of social insurance, including unemployment insurance, in chapter 11.

There is another kind of social insurance that is an important kind of economic right—bankruptcy protection. It may seem strange to think of the right to declare bankruptcy without losing all of one's assets as an important right, but it is. For one thing, it is a protection against a kind of temporary slavery. In the nineteenth century, before there was bankruptcy protection, debtors could be imprisoned and made into temporary slaves until they worked off their debts.

Bankruptcy protection goes against ground-level moral judgments that people should have to pay their debts. In this way it is like the doctrine of efficient breach of contracts, in which a party chooses to pay damages for breach rather than to perform on the contract. Both legal doctrines encourage people to break their promises and default on their debt obligations when it is efficient to do so. Thus, both doctrines are exceptions to ground-level moral norms. In the case of bankruptcy protection, the exception is endorsed by the main principle because equity requires only that bankruptcy be bad enough that people will make great exertions to avoid it, but no worse than that. As we saw in chapter 6, even when ground-level norms would require punishment—for example, of those who break their promises—the main principle can endorse exceptions to the punishment norms when the exceptions equitably promote life prospects.[38]

A Right to Gainful Employment?

In a capitalist economy, life prospects will be bleak without gainful employment. This does not imply that there should be a guarantee of full employment. So long as there are rights to social insurance, including unemployment insurance, to provide a safety net for those who are out of work, no one has a right never to be out of work. I discuss social insurance rights in chapter 11. However, except perhaps in times of severe economic crisis, governments should promote gainful employment indirectly, by providing the economic framework within which people can obtain gainful employment. Understood this way, economic rights should include a right to gainful employment.

Efficiency or Well-Being?

Most of the examples of improvements in the law of property and contracts that I have discussed in this chapter are familiar from the law and economics literature. However, in that literature, the main proposal has not been that improvements in the law should be defined in terms of equitably promoting well-being. Instead, it has been thought that improvements should be defined in terms of promoting efficiency, where efficiency is defined in terms of what exchanges people actually agree to or would agree to, itself determined by what people are willing to pay and what they are willing to accept in voluntary exchanges. There are two commonly employed standards of efficiency in the literature. The first, Pareto efficiency, is the least controversial, because it simply deems inefficient any outcome in which some parties would be willing to enter into voluntary exchanges. Once those voluntary exchanges have taken place and no parties are willing to engage in further exchanges, the outcome is Pareto efficient.

The second criterion of efficiency is parasitic on the first, but in a way that makes it more controversial. To illustrate the difference, consider a variation on the case of Marie, the medical researcher. Marie is the sole producer of a life-saving drug that she sells at her profit-maximizing monopoly price of $1,000 per dose. Suppose that the marginal cost of producing a dose of the drug is $1. That is what the price would be if other producers could obtain the formula for the drug and compete with Marie. If we set aside questions about what is necessary to motivate medical researchers like Marie to do the research necessary to develop new drugs, it is clear that there is a Pareto-superior outcome to the status quo. It is an outcome in which all the current buyers pay $1,000 per dose, so none of them is worse off, but other potential buyers who are not willing (or able) to pay $1,000 per dose pay what they are willing to pay, so long as they are willing to pay at least the marginal cost of a dose, $1. Note that this departure from the status quo can be reached by a series of voluntary transactions. Because each of the transactions would increase Marie's profits, both parties would voluntarily enter into them.

Because a market system typically does not permit differential pricing, a market system would typically not reach this Pareto-superior outcome.

The Kaldor-Hicks criterion of efficiency is a corrective to this kind of Pareto inefficiency. What is controversial about the Kaldor-Hicks criterion of efficiency is that, starting from the fact that the status quo is not Pareto efficient, it can be used to justify a move *not* to the outcome in which differential pricing maximizes the profits of the seller, but to the outcome that maximizes the benefits to the buyers, simply because those benefits are great enough to *potentially compensate* the seller. For this reason, the Kaldor-Hicks test is referred to as a *potential compensation test*.

The Kaldor-Hicks potential compensation test has serious problems, but from the point of view of the main principle, it often gives results that are superior to the Pareto test. For example, the main principle typically endorses antitrust legislation designed to prevent producers from earning monopoly profits, because, evaluated as a policy, such legislation benefits the relatively large body of consumers, including the less well off, while reducing the profits of only a relatively small body of producers, who are among the more well off. Behind the veil of ignorance in the EOP, no one would argue for protecting monopoly profits, unless, as in the case of intellectual property rights, those profits were necessary to motivate investment that would significantly increase life prospects.

So the main principle and the Kaldor-Hicks test support the same laws for monopolies. However, the Pareto test would not give this result, because it would require that monopolists be compensated for their lost monopoly profits. In the above example, it would require differential pricing, so that those who were willing to pay the monopoly price would pay it. This practice would not be endorsed by the main principle, because it would promote inequity.

Although the Kaldor-Hicks test gives the correct result in some cases, the general idea that outcomes can be justified by potential (not even *actual*) compensation could never be endorsed by the main principle or any other remotely adequate moral principle. Such a principle would not only require governments to locate sewage treatment plants, hazardous dumps, and polluting industries in poor rather than wealthy communities, it would not even require the wealthy communities to compensate the poor communities for assuming these costs. So although the Kaldor-Hicks criterion can be a useful criterion for evaluating alternative policies, it could never, by itself, justify one policy over another.[39]

Posner, who applies the Kaldor-Hicks criterion of efficiency to the law makes an important point (1983, 102).[40] Applying the Kaldor-Hicks criterion to the law is different from applying it on a case-by-case basis. In any individual legal case, there will be winners and losers. But if, for example, over time, we each have the same probability of being a winner as a loser, then over time, a practice that satisfies the Kaldor-Hicks criterion would be one in which everyone could expect to be a net winner over time. This is obviously very close to an EOP argument for practices that satisfy Kaldor-Hicks.

Is there an EOP argument for practices that satisfy Kaldor-Hicks? There would be if applying Kaldor-Hicks to social practices would be expected to equitably promote life prospects. But it is obvious that applying Kaldor-Hicks to social practices would not be expected to equitably promote life prospects. The reasons are reasons that also apply to the Pareto standard of efficiency, so it is useful to discuss that standard also.

Both the Pareto and the Kaldor-Hicks standards of efficiency define efficiency in terms of willingness to pay or willingness to accept payment. Because both willingness to pay and willingness to accept are themselves highly dependent on one's level of wealth, both standards are standards for implicitly promoting well-being weighted by one's level of wealth. Under either standard, the life prospects of the better off count more than the life prospects of the less well off. Perhaps some such weighting could be justified if there were some reason to think that the economic system rewarded people with the level of wealth that they *deserve*, so that there was some basis for thinking that the better off *deserved* to have their life prospects weighted more heavily than the life prospects of the less well off. But in an economy in which income is based on commodity value, this seems crazy. Perhaps a moral case can be made for a system in which teachers earn only a fraction of what movie directors earn, but the justification could not be that teachers deserve to have their life prospects given less weight than the life prospects of movie directors.

Of course, if everyone's level of wealth were approximately the same, then there would be no bias against the less well off. This is not an alternative that is available to the advocate of efficiency, because both standards of efficiency, Pareto and Kaldor-Hicks, favor practices that will inevitably generate great disparities in wealth. As we saw in the discussion of commodity value, it is those very disparities that are necessary for efficiency, because they are the signals and rewards necessary to motivate productive investment.

To take only one example, neither the Pareto nor the K-H standard would rule out systems of slavery, because systems of slavery can clearly be efficient based on either standard (Satz 2009). Marie the medical researcher is an example of a system of slavery that is clearly Pareto and K-H efficient, and it might well be that the general practice of permitting such contracts would be Pareto and K-H efficient also. And it is not only hypothetical systems of slavery that pass these efficiency tests. Good arguments have been given for thinking that slavery in the antebellum South was both Pareto and K-H efficient (Fogel and Engelmann 1974). But no one could think that such systems of slavery equitably promote life prospects, and no one would advocate them behind the veil of ignorance in the EOP.

It is hard to deny the moral force of these sorts of equity considerations as objections to standards based on efficiency. Posner himself admitted as much when he acknowledged that there are cases in which one recoils from the implications of allowing markets to control all allocation (1983, vi). His own example is one in which a wealthy person of normal height is permitted to

buy all of a scarce growth hormone to increase his height by a couple of inches, when it could be used to help someone born with dwarfism to reach a normal height. So equity matters.[41]

Willingness to Pay for Equity

Perhaps my argument is too quick. Zerbe has suggested that the kinds of arguments I have given above implicitly assume that people are self-interested and that they are only willing to pay for goods and services that benefit them. The conclusions don't follow if we assume that they are willing to pay for goods and services that benefit other people—in short, if people are willing to pay for equity. Zerbe allows for utility measures that include any values for which people are willing to pay, including distributional values such as equity (2001, 17–18).

If people, especially the more well off, are willing to pay for improvements in equity, then it might well be that the practices that qualify as efficient under Zerbe's expanded Pareto or K-H tests would in fact promote equity.[42] This is an interesting suggestion that deserves serious consideration, but it cannot do the work that theories of justice are designed to do. To see this, notice that Zerbe's suggestion generates a new version of the Euthyphro question:[43] Does justice demand equity because people are willing to pay for equity, or are people willing to pay for equity because it is demanded by justice? To say that justice demands equity because people are willing to pay entails that if people were not willing to pay for it, it would not be a demand of justice. It is easy to imagine a situation in which a majority exploits a minority but has no empathy for the minority and would not be willing to pay anything to promote equity. Such insensitivity could not make the arrangement a just one.

However, if Zerbe's proposal is not interpreted as a proposal about what makes a situation just, but rather as a proposal for understanding how a market system might help to make the world more equitable, it has a lot to recommend it. Although in 2009 the worldwide crisis of market capitalism grabbed all the headlines, the most important development in late twentieth-century and early twenty-first-century capitalism is not this reminder of lessons about the problems of laissez-faire capitalism learned in the past and forgotten, but the way that willingness to pay for equity and other values is transforming capitalist enterprise. I live in Seattle, which is at the forefront of this change, so my experience may not be representative. Where I live, some car dealers don't advertise their cars, they advertise the carbon offsets they have purchased for each vehicle sold; some fast food purveyors don't advertise their food, they advertise the environment-friendly containers that they put it in; some coffee shops don't advertise their coffee, they advertise their fair treatment of the workers who plant and harvest the coffee and their environmentally friendly practices. My orange juice container offers me the

opportunity to save 100 square feet of rain forest. Wal-Mart advertises how well it treats its employees. And when I want to support entrepreneurs in the developing world, I can log onto kiva.com and join with thousands of others who are making microloans to people in need all around the world.[44]

This is a bottom-up movement that has the potential to make capitalism an almost irresistible force for promoting equity. And it all can be traced to what Zerbe has tried to draw our attention to, even though it is invisible in most mainstream economic analysis: that people, including the relatively well off, are willing to pay for improvements in equity. So capitalism has the potential to promote equity and other social values in a way that even Milton Friedman (1970) would endorse.

Conclusion: Economic Rights as Human Rights

In this chapter, I have explained why a suitable package of private property and contract rights based on properly regulated markets and including workers' rights to minimum wage and maximum hours protections, occupational safety and health protections, to gainful employment, to unemployment compensation, and the right to unionize, as well as a general right to bankruptcy protection, would be favored by the main principle as universal human rights. In brief, the reason is that these rights can be used to define an economic system that functions as a self-regulating system for equitably promoting life prospects. In my first volume (Talbott 2005), I explained why the basic human rights were needed to define a political system that would function as a self-regulating system for equitably promoting life prospects. Why are such self-regulating systems so much more effective at equitably promoting life prospects than top-down systems of command and control led by an individual or small group? In the first volume, I identified two problems with such top-down *political* systems, a motivation problem and an information problem. The same two problems are fatal to top-down command and control *economic* systems. The motivation problem is the most salient. If there is a position of power that includes control over the entire economy, the competition for that position will almost guarantee that whoever wins it will use the power to promote his own life prospects rather than to equitably promote the life prospects of everyone.

The information problem is deeper. It is an epistemological problem. To understand it, it is necessary to understand how a market economy can model Mill's social epistemology. It was Hayek (1960) who made the connection explicit by explaining how a market economy not only makes possible but facilitates each of us benefiting from the special knowledge and skills of all other participants in the market. No single individual or small group could possess anywhere near this amount of knowledge, so no top-down command and control economic system could have nearly the potential for promoting the life prospects of those who participate in it. The great challenge in

designing a market system is to design one that will *equitably* promote life prospects. No human being could have solved this design problem on his own, but over hundreds of years, the social process of making gradual improvements through changes in common law and in statutory law has taken us a long way toward solving it.

The idea of an ideal market system as a self-regulating system for equitably promoting life prospects can help to explain what would otherwise seem to be an invidious distinction among disabilities, one that is endorsed by the main principle. Recall that in chapter 4 I claimed that the main principle could endorse a system that compensated the congenitally paralyzed, whose paralysis was due to genetic factors over which no one had any control, at a different level from those who were paralyzed by defective products. This paradoxical result is explained by the fact that the congenitally paralyzed would be compensated from the system of social insurance endorsed by the main principle, which I discuss in chapter 11, whereas those paralyzed because of defective products would be compensated from the income of the manufacturer of the defective product under a system of strict liability for harm. Compensation payments made by the manufacturer would constitute negative feedback in a self-regulating system designed to motivate the manufacturer to shut down production of the product if the compensation costs were so high that selling the product were no longer profitable. As I explain in chapter 11, there is no reason to expect that the level of compensation set by a system of social insurance would be as high as the level of compensation required to make manufacturers appropriately responsive to the damage caused by their products.

Because an ideal market system operates as a self-regulating system for equitably promoting life prospects, economic rights are a microcosm of all human rights, which are the rights necessary for government as a whole to operate as such a system. Economic rights are also a model for human rights in another way. Markets motivate people to make productive investments to improve the goods and services available for purchase. Human rights enable and motivate people to make investments in themselves to improve their lives and give them the best life that they can have.

TEN

Democratic Rights

> The fairness of the rule is a property of the rule itself, and can be established without any need to predict what the outcome of it will be at any particular time and place.
>
> —Brian Barry

> By their fruits shall you know them.
>
> —Jesus (Matthew 7:15–16)

In my first volume (Talbott 2005), I included as basic human rights democratic rights constrained by constitutional protections of autonomy rights enforced by an independent judiciary. I included democratic rights as basic human rights, because of their role, in combination with autonomy rights, in solving what I called the *reliable feedback problem* and the *appropriate responsiveness problem*. The consequentialist version of the former is the problem of governments' obtaining feedback on how well (or poorly) their policies equitably promote life prospects. The consequentialist version of the latter is the problem of making governments appropriately responsive to feedback, so that they tend to continue policies that do a good job of equitably promoting life prospects and to discontinue or modify policies that do not do a good job of it. In this chapter I consider what kind of democratic rights would be favored by the main principle. Although my focus is on democratic rights, it is important to remember that my defense of them depends on a background of constitutional protections of autonomy rights. Without protections of those rights, there is no consequentialist defense of democratic rights. Historically, lots of dictators have used their control over the media and the ability to suppress political opponents to guarantee themselves election by large majorities.

Before taking up the consequentialist case for democracy, it is useful to begin with nonconsequentialist accounts. One of the most important developments in political philosophy has been the development of philosophical theories of deliberative democracy. In this chapter I discuss a number of such theories as one example of a larger category of "ideal procedure" theories of democratic rights. I argue that although such accounts can provide us with ideals that can inspire improvements in our actual democratic practices, they cannot actually explain why the current practices are justified, or why some of the potential improvements to the current practices would be improvements. Indeed, if the ideal procedures are taken as blueprints for how to

improve our democratic institutions in the actual world, the results could be disastrous.

The reason is that democratic practices in the actual world are far from the ideal, and not all changes to current practices that would make them more closely resemble the ideal would equitably promote life prospects. Indeed, some would greatly diminish the life prospects of almost everyone. I argue that what makes the models of deliberative democracy and other ideal procedure theories ideals is not any intrinsic property of the ideal procedure, but rather the *results* of the ideal procedure. When we turn our consideration to real-world democratic procedures, we find that it would be a mistake to model them too closely on the ideal procedures, because in the actual world, attempting to approximate the ideal procedures would, in some cases, produce worse *results*. Of course, it is open to the nonconsequentialist to explain the badness of the results in nonconsequentialist terms—for example, in terms of equal consideration or fairness or hypothetical consent, and, indeed, for reasons that I have already discussed, especially in chapter 5, the main principle endorses our evaluating democratic institutions in these terms. What I try to show is that those terms of evaluation are an indirect way of producing changes in institutions that tend to equitably promote life prospects.

Although not all attempts to approximate the ideal procedures would improve actual democratic procedures, these ideal procedures of deliberative democracy have the potential to generate ideas for new institutions that would be improvements over the status quo. One such institution that has emerged over the past 20 years is deliberative polling (Fishkin 1991). In this chapter, I explain that proposal and extend it to outline a radical proposal for altering democratic rights, election by deliberative poll, that, were there not a potential for abuse, might very well be a substantial improvement over the status quo, when evaluated under the main principle.

Some Nonconsequentialist Accounts of Democratic Rights

The nonconsequentialist rationales for democratic rights that I address fall into four broad, and sometimes overlapping, categories: (1) ideal procedure derivations of democratic rights (e.g., Barry 1995; Rawls 1993); (2) ideal procedure approximation rationales for democratic rights (e.g., Habermas 1996); (3) other ideal standards (e.g., Kant); and (4) nonconsequentialist arguments for a particular democratic decision rule, typically majority rule (e.g., Elster 1993 and Waldron 1998 and 2006). I discuss them separately.

1. Ideal Procedure Derivations of Democratic Rights

Rawls (1993) is a representative of an ideal procedure derivation, because he claims that in the ideal procedure of the original position there would be unanimous agreement on democratic rights in a constitutional framework.

Barry (1995) employs a Scanlonian rather than a Rawlsian original position to argue for a majoritarian democracy in a constitutional framework. There are three problems with this sort of ideal procedure derivation of democratic rights.

The first problem is that such accounts must address a version of the Euthyphro question: Is the hypothetical agreement just because it is the result of the relevant process (the *pure procedural* answer, cf. Rawls 1971, 85–86) or is it the result of the relevant process because it is just (the *tracking* answer)? If the process is a process of hypothetical consent based on rational/reasonable agreement, then it seems that that the account is committed to there being something that makes the agreement rational/reasonable, and thus the procedure would seem to be tracking something independent of the procedure (i.e., the tracking answer). What might it be that makes the hypothetical agreement rational/reasonable? As I discussed in chapter 2, it is unavailing to try to answer that question in terms of hypothetical agreement. I tried to answer this question for Rawls's theory in chapter 4. Recall that Rawls himself thought that the general terms of agreement would be agreement on his general conception of justice, given by the maximin expectation principle. This suggested to me that what underlies the reasonableness of the agreement in Rawls's original position is some conception of the equitable promotion of well-being.

Barry's Scanlonian test requires agreement based on a process that excludes proposals that anyone could reasonably reject.[1] So his theory also tracks something—namely, whatever it is that makes proposals not reasonably rejectable—and thus falls under the tracking answer to the Euthyphro question also. Scanlon argues that consequentialism can't explain the grounds for reasonable rejection. It is quite clear that utilitarianism can't capture Scanlon's conception of reasonable, because it is easy to think of examples in which someone can reasonably reject an action that would satisfy the utilitarian principle (e.g., Scanlon 1998, 235). Scanlon also correctly points out that the variety of reasonable considerations do not all have to do with well-being (chapter 3). This shows that no ground-level consequentialist principle can capture all the grounds for reasonable rejection. However, it does not rule out a meta-level consequentialist explanation. In chapter 9, I proposed a meta-level consequentialist explanation of one of Scanlon's paradigmatic nonconsequentialist reasons, the noninstrumental value of consent. There I showed that the fact that considerations stated in terms of the noninstrumental value of consent do not mention well-being is compatible with there being a meta-level consequentialist explanation of the force of those considerations. A similar argument could be made for his other examples, such as friendship and family relations (174), sexual relations (175), promises, and the principle of fidelity (chapter 7). So Scanlon has not ruled out a consequentialist meta-level explanation of reasonableness.

I conclude that the ideal procedure derivations that give the tracking answer to the Euthyphro question do not rule out a meta-level consequentialist explanation in terms of equitably promoting well-being.

What about the other answer to the Euthyphro question, the pure procedural answer? Does it make sense to think that any procedure could guarantee just results? How could we ever believe that the results of a given procedure could determine what justice (or morality or legitimacy) requires unless, at the very least, we had some way of assuring ourselves that its results would not be awful? No procedure could make it just (or right) to torture babies for the fun of it. But even to acknowledge that there is some moral constraint on the results of the relevant procedure requires us to give up the pure proceduralist answer to the Euthyphro question.

The pure procedural answer faces another challenge also: If there were a procedure that determined the requirements of justice (or morality or legitimacy), we would have to have *a priori* insight into what procedure it is. If we had *a priori* insight into it, all rational beings would agree on it. This can't be true because it is hard to get any kind of agreement on an ideal procedure for justice (or morality or legitimacy). Rawls was right. If there is some ideal procedure, the only way we could figure out what it is would be by a reflective equilibrium procedure that crucially involves making judgments about the acceptability of its results (1971, 21, 141). If there is an ideal procedure, the only way we would have of figuring out what it is would be to be able to find a real-world procedure that produced results that were generally regarded as just. Then we could project the features of the ideal procedure from our real-world procedure.

A second problem with ideal procedure derivations of democratic rights (or of other rights) is that the rights that they produce are rights for an ideally just society. This leads to a further problem that is a problem for all ideal procedure theories.

A General Problem for Ideal Theories. *Ideal* theories of justice or legitimacy are theories of justice or legitimacy for a fully just or fully legitimate civil society. Ideal theories of justice or legitimacy need not invoke any ideal procedures, but, as illustrated by Rawls, they often do. Rawls invokes the procedure of the original position to derive the principles of justice for the background institutions of an ideally just society.

All such theories have the following problem: How do they apply to a less than ideal society such as ours? A natural suggestion is that they provide a comparative standard for improving less than ideal societies—to make them more closely approximate the ideal. If the ideal standard is consequentialist, then the comparison will be based on results. But if the ideal standard is nonconsequentialist, then the comparison will be based, at least in part, on something other than results. In this section, I show how standards of improvement not based on results have unacceptable consequences.

I have already illustrated the problem for ideal theories in my discussion in chapter 7 of the accounts of rights to freedom of expression of both Rawls and Habermas. Neither account would support a right to freedom to express intolerant subversive advocacy, because in an ideally just society no one

would claim a right that she would not be willing to grant to others. But in the less than ideally just actual world, a right to freedom of expression that covers intolerant subversive advocacy is an improvement over the ideal version of the right that does not cover it.

The same problem arises for a theory like Barry's, based on a Scanlonian original position. In the Scanlonian original position, in which everyone is seeking agreement on terms that no one could reasonably reject, no one would propose a right to intolerant subversive advocacy, but if, somehow, such a right were placed on the agenda, everyone would reasonably reject it on the grounds that those proposing it were asserting a right that they would not be willing to grant to others. This kind of consistency is Barry's main test of impartiality (1995, 83–84).

The example of a right to intolerant subversive advocacy is a simple example of a much larger problem for ideal theories. In a nonideal world, it is often better not to try to approximate the requirements of an ideal theory.

2. Democratic Rights as an Approximation of an Ideal Procedure: Habermas

I have already discussed Habermas's theory in broad outline in chapter 7. Here I focus on more of the details of the process of ideal rational discourse.[2] Habermas holds that the legitimacy of laws depends on the *process* by which they are adopted, where the standard of legitimacy is the ideal of rational discourse, discussed in chapter 7. Positive law can derive its legitimacy only from "a procedure of presumptively rational will formation" (1996, 457). Democracy is the form the discourse principle takes in the enactment of positive law (455). As he says, "The *democratic process* bears the entire burden of legitimation" (450). For the democratic process to sustain this burden, it must produce rational outcomes in a procedural sense—that is, as closely enough approximating the ideals of rational discourse (453).

I have already mentioned an epistemological problem with this view. How could we know that any real-world democratic process closely enough approximates the ideal process to legitimize the results of the democratic process? Habermas insists that there is no way to determine the results of the ideal process other than by letting it run, so there is no way to defend a real-world procedure by comparing its *results* with the ideal procedure, because we can, by definition, run only actual procedures, not the ideal procedure.

Perhaps Habermas's theory could at least provide us with a standard of comparative legitimacy: The *more closely* a real-world process approximates the ideal process, the *more legitimate* its laws. If we took the comparative standard seriously, we would have to compare existing democratic processes with the process of ideal rational discourse. The defining characteristic of ideal rational discourse is that it is *not* a competitive, strategic interaction, in which one tries to obtain an agreement most favorable to oneself. It is a cooperative interaction in which *everyone affected* participates, with the goal of

freely reaching a unanimous agreement on what is best supported by the reasons (Habermas 1990, 58–59, 66, 88–89).

No existing democratic process approximates this ideal very closely. This is not necessarily a problem for Habermas's account, because it would be open to Habermas to respond that no actual democratic system has much legitimacy. A more serious problem for his account is that we can recognize that there are changes that would make existing democratic systems more closely approximate the ideal of rational discourse, but to make those changes would potentially make the system worse, not better. Consider three potential changes:

(1) *Direct democracy*. Existing democratic systems are representative democracies. Only elected representatives are allowed to speak in legislative debates. It is true that everyone has freedom of speech to discuss the issues and Habermas places great weight on this as essential to democratic legitimacy, but it must be admitted that the democratic *process* would more closely approximate the ideal if everyone were permitted to participate. There was a time when it was possible to hold that the only feasible approximation of the ideal was a representative democracy. This is no longer true. It would be possible to use the Internet to allow everyone to participate in discussing and voting on legislation. It is quite possible that such a process would produce a hodgepodge of laws that did not form a coherent whole. If so, I take it that no matter how much the process resembled the ideal process, the results would favor representative democracy. Of course, for Habermas, results cannot enter into the evaluation.

(2) *Unanimity rather than majority rule*. No democratic legislature operates by consensus—that is, no democratic legislature requires unanimity (or even near unanimity) to enact a law. Obviously, there are no practical impediments to adopting a unanimity rule. If anything, the case for a unanimity rule is stronger in a representative democracy than in a direct democracy. Because each legislator may represent a million or more people, a single nay vote can represent one million individuals who do not agree with the legislation. Clearly, a process that allows an outcome that one million voters would object to does not come close to approximating Habermas's ideal.

Habermas endorses Fröbel's response to this problem, which is that, because "laws require the justified consent of all" (1996, 475), we must understand the legitimacy of majority rule in terms of a "*conditional* consensus," a consensus to be bound by the will of the majority (475). The only problem with this conclusion is that no constitution establishing majority rule has ever been adopted unanimously by the citizens of any democracy, so what procedural reason could there be for thinking that an ideal process of rational discourse would produce unanimous agreement on majority rule? I discuss substantive reasons for favoring majority rule shortly.

(3) *Strategic interactions*. No observer of existing democracies would ever suffer under the misapprehension that political will formation in democracies occurs by a cooperative process of reasoning in which each legislator

seeks terms that could freely be agreed to by all. Democratic politics operates by coalitions that attempt to force minorities to accept laws that they strenuously object to. Alliances are often built on quid pro quo exchanges and other strategic considerations.

Habermas recognizes this fact, and so he emphasizes that the process of discourse includes not only legislative debates, but also the open discussions among citizens in the public sphere.[3] However, even the discussion in the public sphere is often only distantly related, if at all, to discussion aimed at reaching a conclusion that could be the object of unanimous unforced agreement. One potential "improvement," to make the existing process more closely approximate the ideal, would be to legally prohibit expression in the public sphere that appealed to any reason that could not be accepted by everyone. Not even Rawls, who articulates such a *moral duty* of public reason on constitutional essentials and matters of basic justice, would support legislating that duty (1993, 217). Though a good case can be made that such a law would make existing democracies more closely approximate the process of ideal discourse, I think it is clear that such a law would have disastrous results.

My conclusion is that even as a standard of comparative legitimacy, Habermas's ideal procedure approximation standard fails. If the closest approximation of rational discourse in this world would have all sorts of bad effects, then it is better for those of us in this world to have a system that is not such a good approximation of the ideal. Of course, even to make such a judgment requires us to be able to evaluate processes by their results.

More on Rawls. Although much of Rawls's theory is an extreme idealization, his account of democratic rights seems much more modest and down to earth. All he requires is that we guarantee each person the fair (not even equal) value of the political liberties. Is this modest requirement such an idealization that the ideal theory objection applies to it?

Rawls himself was not under any illusion that the fair value of the political liberties is guaranteed in the United States, because of the role of money in electoral campaigns (1993, 356–363). Half the members of the U.S. Senate are millionaires. Of course, not all of them were millionaires when they arrived, but most of them were. What percentage of them had blue-collar jobs before entering politics? The House of Representatives is the house of the people, but over a quarter of its members are millionaires, too.[4] No one could think that the United States assures fair value of the political liberties.

With public funding of campaigns it would be possible to make the value of the political liberties less unfair, but it is hard to think that real-world democracies should be willing to spend whatever it takes to make the value of the political liberties truly fair. In the actual world, the most that we can reasonably aspire to is that their value be not too unfair.

How unfair is too unfair? There are two ways of trying to answer this question—one procedural and one based on results. As I explain in the

remainder of this chapter, I think the only remotely plausible approach is one based on results. What would a procedural approach be like? The procedural approach would require comparing the existing system with an ideally fair system and somehow measuring how close the existing system comes to the ideal system. Then it would be necessary to have some principled way of setting a limit to how far from the ideal counts as too far. I think the most plausible procedural approach to evaluating the value of the democratic rights would take some kind of majoritarian system as ideal, so I postpone my discussion of this objection until I discuss majority rule.

3. Other Ideal Standards

The fair value of the political liberties is only one among a family of abstract nonconsequentialist standards—for example, equal respect, equal dignity, equal concern, and equal consideration. For each standard, the question is whether we should evaluate the adequacy of the status quo on procedural grounds alone, or whether we should do so on the basis of results. Each of these standards can be given a procedural and a results-based interpretation. The question is whether a procedural interpretation is adequate.

Consider first standards based on equal dignity or equal respect. Respect is usually thought of as the appropriate response to dignity (Darwall 2006, 119), so I discuss respect, with the understanding that parallel considerations apply to dignity.

To be at all plausible, a standard of equal respect has to apply to recognition respect, not appraisal respect. Appraisal respect is respect we earn; recognition respect is the respect we are owed as persons (Darwall 2006, 122–124). So the moral standard we are interested in is equal *recognition* respect.

To the nonconsequentialist, recognition respect just is what is called for by the dignity of persons. On the Kantian [1785] account, it is called for by the recognition that each person has incomparable worth. But here we confront a puzzle. No one can seriously think that this is really true. If each person has incomparable worth, then we would be acting irrationally if we ever chose an act that had a higher probability of leading to a human death over an act that had a lower probability of leading to a human death. But we make such choices all the time—for example, by driving in a car—and it would be irrational not to.

Perhaps Kant made a mistake. Perhaps equal respect does not require that everyone have incomparable worth, but only that they have the same finite, equal worth. But this can't be right either, because if the worth of each were finite, then there would be a fair price for selling oneself into slavery or for selling to someone else the right to kill you.

One possibility for understanding equal respect is in terms of some sort of equality of status. This connects with the other ideals mentioned above. The suggestion would be that equal respect is to be understood in terms of equal

concern or equal consideration. But no political system gives even remotely equal concern or consideration to every citizen. Even if the U.S. political system were not so distorted by the influence of campaign contributions, there would be no plausible case for thinking that the rich and the poor receive anything like equal concern or equal consideration.

So these standards generate the same problem as Rawls's standard of fair value. How unfair or how unequal is too unfair or too unequal? Again, there are two ways of answering this question, one procedural and one results-based. I illustrate the general problem for all procedural answers by considering in some detail the most plausible procedural account of fairness or equality: majoritarian democracy based on one person, one vote.

4. Nonconsequentialist Arguments for a Particular Democratic Decision Rule, Typically Majority Rule

Waldron (1998 and 2006) favors majority rule on grounds of equality and fairness. Waldron opposes any limitations on majoritarian rights. This does not rule out constitutional limitations (so long as the constitutional limitations are adopted by a majority), but it does rule out judicial review of majoritarian legislation. In this section I discuss whether there is some special property of majority rule that makes it intrinsically fair, or whether its fairness depends on results. I begin with a particularly elegant and unqualified version of a procedure-based rather than results-based argument for majority rule from Elster:

> Although one may believe that majority rule needs to be limited and constrained in various ways, these limits and constraints can ultimately have no other normative foundation than a simple majority decision. Consider the ideal case of a constituent assembly operating in a complete historical and social vacuum, for example, a group of settlers writing a constitution for their new country. Although the assembly may decide that a qualified majority shall be required to change the constitution, that decision itself must be taken by a simple majority. If one required a qualified majority at the constitutional convention, two problems arise. First, the assembly might not be able to produce a constitution at all. In constitutional amendments, the existing document serves as the status quo that remains in force when a proposed amendment fails, but in a creation *ex nihilo* there is no status quo that can serve as a fallback position. Second, and more important, the decision to use a qualified majority would itself have to be made by a simple majority, to avoid an infinite regress. (Elster 1993, 179–180; footnotes omitted)

Elster claims that there can be no other normative foundation for any group decision rule than majority rule. This is a very strong claim. What is

surprising is that he thinks he could establish such a claim by what seems to be an *a priori* argument about the logic of majoritarianism. How could this be? How could a logical truth determine a normative truth about legal legitimacy?

It is important to emphasize that Elster is making a normative claim. Elster's two arguments for it are very brief. With respect to the first argument, why should he suppose that the *fact* that in some situations no constitution would be adopted unless majority rule were employed settles the normative issue? Even when settlers are writing a constitution for a new country, it would seem possible that no constitution at all could be morally preferable to some constitutions that would be approved by a majority. I suspect he may be thinking that it would be objectionable in the constitutional case for a minority to be able to block a majority decision. But how could that be decided *a priori*? Certainly, if the constitution that the majority approved legalized the enslavement of a minority by a majority, the fact that the majority approved of it would not justify it.

Elster regards the second argument as the most important one. But the second argument is even more puzzling than the first. It seems to amount to the claim that a majority would favor majority rule. But if majority rule is not morally justified, the fact that a majority favors it won't justify it. Elster seems to assume that nothing could justify a decision rule except the application of a decision rule. If so, majority rule is the only rule that can be pretty much guaranteed to justify itself. But there is an alternative—that what justifies a group decision rule is *something other* than a group decision rule.

Perhaps Elster is thinking that there is an epistemic problem lurking here. Suppose, for example, that there is universal agreement that a group decision rule must be justified by something, but there is disagreement on what that is. Suppose a majority agrees that to be acceptable, a group decision rule must satisfy acceptability test A. Elster seems to be suggesting that that fact alone would be sufficient to justify making A the acceptability test for a decision rule for the group. There is something to be said for this result, because the alternative would seem to require giving extra epistemic weight to the opinions of a minority, rather than giving every opinion the same epistemic weight. But not every opinion *should* be given the same epistemic weight. For example, acceptability test A might be that the preferred decision rule best promote the interests of white citizens. If whites were in the majority, they might well have self-serving reasons to agree on A as an acceptability test for a group decision rule. But their agreement would not justify making A the acceptability test for their group decision rule.[5]

How could Elster have thought that we could draw normative conclusions about a decision rule from its logical properties? Well, there are some normative conclusions that we can draw about the nature of majority rule. I illustrated one of them in my discussion of the coordination problem in chapter 3. When the main principle favors a randomizing procedure, in order to implement the procedure, it is necessary to settle on a single application of a

single randomizing device. If three of us are in a situation in which tossing a coin is a fair way to decide an issue and each of us has a coin to toss, there must be a way of determining which coin and which toss are decisive. In cases such as this, majority agreement on which coin and which toss are enough to resolve the indeterminacy.[6]

But Elster is not talking about solving a coordination problem among equivalent alternatives. He is talking about a choice among nonequivalent alternatives, such as constitutions. How could the formal properties of a decision rule settle the question of its appropriateness for choices in which what is chosen matters? Elster could not really think that a majority could justify its enslavement of a minority by a majority vote on a constitution that made their enslavement legal.

All procedural defenses of majority rule face the same kind of problem: the possibility of majority approval of legislation that has very bad results for a minority. This suggests that proposed improvements to majority rule should focus on results—on making the results of the rule such it does a better job of equitably promoting life prospects. But there is no procedure that can be guaranteed to equitably promote life prospects. A procedure that is adequate in one context may be awful in another context. Whatever our standard for evaluating improvements in a political system—whether it be equal respect or equal dignity or equal concern or equal consideration—any adequate test for how to get closer to the standard should consider results, especially the consequences for the equitable promotion of life prospects. How could we be morally required to comply with majority rule or any other procedure if it produced awful results?

How could we know that eliminating judicial review, as Waldron (2006) suggests, would be a moral improvement? It is true that the U.S. Supreme Court has historically favored the interests of the very well off over the interests of the less well off. That is a results-based consideration that would have to be balanced against cases in which the Court's decisions have improved the life prospects of the less well off. A procedural defense of majority rule would somehow avoid such balancing.

It is true that the main principle does not endorse our applying it directly to judge improvements. It favors ground-level norms of equal consideration or equal respect or equal dignity. But in applying those ground-level norms, we cannot limit ourselves to the formal features of decision rules; we need to be sensitive to their results.[7]

An Alternative to Democratic Rights? Election by Deliberative Poll

Once we acknowledge that the moral case for democratic rights depends on results, we must acknowledge the possibility that there might be an alternative that would be a moral improvement. In this section I discuss an

alternative suggested by the ideal models of deliberative democracy discussed earlier—an alternative based on Fishkin's (1991) use of deliberative polling.[8] Fishkin has championed deliberative polls as superior to standard polls on political questions. I extend Fishkin's proposal and consider whether deliberative polls might also be an improvement on democratic elections. Briefly, the idea would be to select a statistically representative sample of the U.S. population, large enough to provide a good cross section of the country but small enough to bring them together for a sustained period (e.g. a week) of education, candidate forums, and group deliberations. At the end of the week, this sample of the American population would elect the president. Call this *election by deliberative poll*.

Such a procedure would solve a number of problems with our current electoral system. One big problem with the current system is that candidates are required to raise large amounts of money to conduct a campaign. This gives large donors an undue influence over the government and gives wealthy candidates an advantage over poor ones.[9] Election by a statistically representative deliberative poll could end this influence.

Another big problem with the current system is that those who actually vote in elections are not representative of the general population. The more educated are more likely to vote than the less educated; the more affluent are more likely to vote than the less affluent; the older are more likely to vote than the younger; whites are more likely to vote than nonwhites (though this may simply reflect differences in education).

Yet another problem with the current system is that many people have little incentive to become informed voters, because each person's vote has such negligible effect. If a group of, say, 5,000 voters determined the results of a presidential election, each vote would be extremely important.

Unfortunately, it seems to me that, in spite of its advantages, election by a statistically representative deliberative poll would be too liable to abuse to be workable. Political parties and special interests would have huge incentives to try to influence the voters in the representative sample. It would be naive to think that they would not figure out a way to do so.

However, let's set aside the potential for abuse to ask whether the procedure itself somehow fails to respect each individual by disenfranchising all but a group of 5,000. What exactly would be the problem? Well, it does seem unfair if I look at it from my particular point of view. It is very unlikely that a representative sample of 5,000 members of the electorate would include even one philosopher. Thus, not only would I not have a vote, there would be no one whom I could think of as representing me who had a vote. It would seem that I am completely disenfranchised. How could this be fair to me?

And yet, if presidential elections were based on informed debate and dialogue among the members of a statistically representative sample of the less affluent as well as the more affluent, the less educated as well as the more educated, young as well as old, nonwhites as well as whites, this would almost surely raise the level of presidential campaigning and produce

presidents and policies that would do a better job of equitably promoting the life prospects of the citizenry.

Suppose we lived in such a system. To take just one example, it is inconceivable to me that voters in such a system would have permitted the government to enact the Bush administration tax cuts, tax cuts that could be projected to generate the large and potentially endless deficits in the government's general accounts, deficits that are routinely underreported to a poorly informed electorate because the official reports allow offsets to the general accounts deficit equal to the amount of the collections for a nonexistent trust fund for the Social Security system. For another example, I doubt that the voters in such a system would have permitted the government to maintain a Social Security and Medicare system that will soon generate huge current account deficits. Consider one more example. The Bush administration did everything it could to downplay the threat of climate change. However, the scientific evidence is so overwhelming that it is very likely that, when presented with the evidence on both sides, a statistically representative sample of voters would have insisted on substantive policies to reduce emissions of greenhouse gases.

Suppose we lived in a world in which election by deliberative poll were in force and, as a result, the government's current accounts budget and the Social Security and Medicare budgets were in balance and the government had taken serious steps years ago to substantially cut greenhouse emissions. Would fairness or respect for individuals require us to replace it with a system in which many voters have little incentive to become informed about the candidates' policies and the results disproportionately reflect the preferences of older, more educated, more affluent, whites?

Election by deliberative poll would eliminate one of the few opportunities that most citizens have to exercise a civic duty. Would this reduce civic virtue? It is difficult to think that replacing universal suffrage with election by deliberative poll would have a significant deleterious effect on civic virtue, because many jurisdictions have compulsory mail-in ballots, and it is just not plausible that voting by mail-in ballot generates much civic virtue. Of course, involvement in democratic processes has some consequentialist value. Call this contribution to well-being the *consequentialist benefit of participating in elections*. On a consequentialist account, election by deliberative poll could be justified only if it generated benefits that outweighed the loss of the consequentialist benefits of civic engagement. It seems to me that the consequentialist benefits of election by deliberative poll might easily outweigh the loss of the consequentialist benefits of civic engagement, as it would, for example, if it had prevented the government from generating the kinds of deficits in its general accounts and its Social Security and Medicare programs that will leave mountains of debt for future generations or if it had motivated the government to adopt serious policies to avert climate change years ago. Could the civic virtues have a nonconsequentialist value that outweighed these consequentialist values? I don't see how.

Unfortunately, I cannot actually advocate a change to election by deliberative poll, because there would be too much potential for abuse. By the standards of current campaign funding, each vote in a deliberative poll for president would be worth more than $50,000 to each candidate. That would pay for a lot of investigation into the lives of the voters and lots of ways of trying to influence their decisions. However, this example still serves as a useful illustration of the fact that, even when, on balance, consequentialist and nonconsequentialist accounts favor the same process (e.g., democratic voting), it is still possible to adjudicate between them. If there were no potential for abuse, would you favor a change to election by deliberative poll?

Rights of Cultural Minorities

My list of human rights has no group rights, only individual rights. Recent human rights documents have included group rights—for example, the African Charter on Human Rights and Peoples' Rights (1986). These rights are often invoked as protections for a minority native population against a majority of nonnatives. Why are there no rights protecting these minorities on my list?

The reason is that I favor a different way of thinking about such rights. There are three ways in which my way of thinking about them differs from the standard model. First, I think of them as the individual rights of the members of the group, not the rights of the group itself. If at some time in the future, none of the members of the group wants to continue the group's cultural practices, I don't believe that the culture itself would have a right to be preserved. If the cultural leaders decided to offer payments to their members to keep them from abandoning the culture, the larger society would have no obligation to fund the payments as part of an obligation to preserve the culture. The minority rights at issue are rights of the members of the minority culture to be able to continue their cultural practices if they want to.[10]

Second, I think of these rights of minorities not as an independent kind of right, but as a limitation on majoritarian democratic rights to assure that those rights equitably promote life prospects. The loss of one's culture and heritage can have catastrophic effects on the life prospects of a people and their descendants. Because the main principle evaluates democratic rights on the basis of their contribution to equitably promoting life prospects, a majoritarian, one-person, one-vote system would not be endorsed if it could be used to eliminate or greatly impair a minority culture. We are already familiar with the need to constrain majoritarian one-person, one-vote with other individual rights. Rights of minorities would simply be another kind of constraint on majorities.

Third, like Kymlicka (1989), I do not see these rights as inherent rights of the members of minority cultural groups, but rather a necessary evil to prevent a worse evil. If a minority culture were not at risk of being dominated by

the majority culture, no such rights would be necessary. Of course, treaty rights and other legal agreements with the dominant culture would be enforced, but that does not require any special kind of right.

The reason I do not think of these rights as inherent is that these rights are often used by minority cultural groups to deny some of the human rights of their members. For example, most native populations exclude women from decision making and give fathers and husbands rights over them that conflict with their human rights. I see no reason to think that male members of native groups have an inherent right to rule over their wives, just because it is part of their cultural tradition. If we make an exception to these human rights to protect the cultural tradition, the hope is that the culture itself will gradually come to respect the human rights of its members, not that it will continue to ignore them.

Though minority cultural rights are typically limited to native populations, the general phenomenon of which they are an instance is not. They are an instance of the need to protect minorities against majority oppression. It is not necessary to be a minority culture to have a need for such protection. In the United States, the assignment of two senators to each state, independent of population, is a departure from one-person, one-vote designed to prevent inequity to the smaller states. If there ever is a global democracy, almost surely some special rights of this kind would have to be defined to protect the less populous nations from being dominated by the more populous ones. They are not inherent rights, because it is possible to imagine that one day the people of the world would have become so cosmopolitan that such rights would have become nothing more than a source of unfair advantage for the less populous nations, as the provision for two senators from each state regardless of population is now an anachronism that continues to be a source of unfair advantage for less populous states in the United States. Such rights should terminate when there is no longer the kind of danger that they are designed to protect against.

How Democratic Rights Promote Life Prospects

On my account, though governments do not and should not always aim at it, there is only one standard for improvements to a legal code: to equitably promote life prospects. In theory, there are two ways that governments promote life prospects: by legal paternalism and by legal solutions to collective action problems (CAPs). I discuss legal paternalism in chapters 12 and 13. Because legal solutions to CAPs don't always promote life prospects equitably, the government should also engage in redistributive practices (e.g., the negative income tax, discussed in chapter 9). I discuss more redistributive practices to promote equity in the next chapter. Here I discuss legal solutions to CAPs.

Legal Solutions to Collective Action Problems

If the government tried to solve only problems that were CAPs for the entire citizenry, it would do very little. If it tried to solve *all* groups' CAPs, it would not do a very good job of equitably promoting life prospects, because it would have to encourage businesses to collude to fix prices or to control supply, it would have to help short sellers coordinate their sales to bring down the value of a stock, to help spammers to coordinate their attacks on computer systems, and much else that would greatly diminish most people's life prospects. The government should solve the combination of CAPs that, taken together, do the best job of equitably promoting life prospects.

There are a seemingly endless number of CAPs. Effective enforcement of human rights is a CAP, as is raising the taxes to fund a government to solve CAPs. No democratic government has ever funded itself with voluntary donations.

Effective security rights are a solution to the internal security CAP. A military for national defense is a solution to the external security CAP. Establishing property and contract rights and markets is a solution to the productivity CAP. As previously discussed, prohibitions on slavery and indentured servitude are solutions to CAPs, as are minimum wage laws, occupational health and safety laws, and collective bargaining laws.

Other solutions to CAPs include the following: legal tender laws; truth-in-labeling laws and mandatory disclosure laws; product safety laws, including laws requiring the testing of drugs and other potentially hazardous products; traffic control laws; licensing laws, antipollution laws; zoning laws; building codes; occupational safety laws; securities laws; antitrust laws; and government investment in medical and other scientific research, in streets and highways and mass transit, in sewers and utilities, in public radio and TV, and in parks and other protected areas. Common pool resources, such as fisheries, also often have the logic of a CAP. Enforceable quotas are a solution.

Not all CAPs are solved by government. Moral systems are solutions to CAPs that are enforced by communities when they are not enacted into law. Communities also develop customary norms for hosting, gift giving, etiquette, and a variety of other matters that make social life go smoothly. These customary norms are typically enforced by social approval and disapproval, not by laws.

It is important to recognize that government solutions to CAPs are not paternalistic. A paternalistic law—for example, a motorcycle helmet law—aims to promote people's good by forcing individuals to do something that they don't think is good for them. It *overrules* their own judgment about what is good for them. In contrast, a legal solution to a CAP forces the members of a given group to act in a way that, when everyone (or almost everyone) in the group acts in that way, they all (or almost all) agree that the law is good for them. It *gives effect* to their judgment of what is good for them. In a CAP, each individual knows that it would be better if all cooperated, but that is not

enough to guarantee cooperation if there is a payoff to defecting. For example, pollution control devices on cars have greatly reduced air pollution in U.S. cities. Probably most people would be willing to pay the cost of a pollution control device for the reduction in air pollution produced by everyone's buying one. But if pollution control devices were voluntary rather than mandatory, very few individuals would be willing to buy one for their cars, because each device would reduce air pollution by only a negligible amount.

Governments must be selective in the CAPs they provide legal solutions to. For one thing, they need to pay attention to all the various costs of enacting and enforcing a solution. They also need to consider how the legal solutions to CAPs fit together into a coherent whole. Finally, they must always be aware that solutions to CAPs almost inevitably beget more CAPs. For example, solutions that require a government bureaucracy generate principal-agent problems—that is, problems of how to motivate the government officials to act to achieve the purpose of the law rather than their own purposes. For another example, legal enforcement of economic rights makes markets possible, but they also make possible Ponzi schemes and bank runs, each of which requires another legal solution. However, when there are constitutional rights to protect minorities against majority tyranny, it is quite plausible that a policy of legally enacting solutions to CAPs that are favored by a majority will, over time, promote everyone's life prospects. When the equity rights discussed in the next chapter and the other human rights on my list are guaranteed, the main principle will favor a majoritarian procedure for enacting legal solutions to CAPs.

Inalienable Rights

In chapter 9, I explained why rights against slavery and indentured servitude are solutions to CAPs. To be effective solutions, the rights must be inalienable. This is the rationale for many, but not all, inalienable rights. Some should be inalienable to promote equity, as I explain in chapter 11. Here I consider the rights whose inalienability is a solution to a CAP.

Political Rawls (1993) offered a nonconsequentialist account of inalienability, but his account is problematic. He held that citizens should be free to agree to limits on their liberties, as in religious vows of poverty or obedience (yet another form of slavery contract), but that such agreements should have no legal force (1993, 365). This view gives the right result for religious vows, but it fails as a general account of inalienability. Consider, for example, vote buying. No vote buyer has ever tried to obtain legal enforcement of the sales agreement. Vote buyers are quite happy if they are permitted to make private transactions not backed up by the enforcement power of the state. But such agreements generate a classic CAP, because voting, at least in large elections, itself typically generates a CAP. It may well be better for each individual to

be able to sell his/her vote, though better for all potential sellers if no one is permitted to sell.

Vote selling is just one example of how, on my account, other people's rights can be more important to my well-being than my own. The same analysis applies to all of the basic human rights, the autonomy rights as well as the political rights. On a simple consequentialist account, it is the value of my exercise of my autonomy rights and my democratic rights that explains their contribution to my well-being. However, on the more complex account that I favor, it is largely other people's exercise of those rights that contributes to my well-being. Indeed, as I understand it, autonomy and rationality themselves are collective achievements. My rationality is a product of the freedom of others, as well as myself, to question authorities, and my autonomy is a product of the freedom of others, as well as myself, to conduct experiments in living and of the ability to learn from the experiments of others, the successful and the unsuccessful. For most people, it would be rational to sell at least most of their autonomy and democratic rights for a relatively small price, because most of the benefits of autonomy and democratic rights derive from *other people's* exercise of them. Thus, autonomy and political rights should be inalienable. Making them inalienable is a solution to a CAP. This is also the rationale for making privacy rights inalienable, as I discuss in chapter 13.

Why Democracies Are Superior to Other Forms of Government

There is no one ideal form of democratic government. Different institutions and decision rules will be appropriate in different contexts. It is possible for the main principle to favor a purely majoritarian democracy in some contexts, but very unlikely in the actual world, because of the potential of tyranny of a majority. Tyranny of a majority is not necessarily a problem for a utilitarian, because if the majority is large enough, its happiness can outweigh the unhappiness of the tyrannized minority. Tyranny of a majority is a problem for obtaining the endorsement of the main principle, because, under the main principle, the distribution of well-being matters and extra weight is given to the life prospects of the less well off.

On an indirect consequentialist account like mine, one must begin by acknowledging the great value of ground-level norms such as one-person, one-vote. A norm like that one, that has played such an important role in the equitable promotion of life prospects, should not be overturned easily. However, in a theoretical discussion like this one, we can recognize that it could be overturned and we can speculate about what norms and practices might be improvements.

The example of election by deliberative poll illustrates one possible improvement. Because of the potential for abuse, it would probably not be an

improvement. Mill made a different suggestion. He suggested the possibility of granting some citizens multiple votes based on their level of education. This seems to me unwise, because it would give more weight to the interests of the more educated than the less educated. Because the less educated would be expected to be less well off than the more educated, it would reverse the weighting favored by the main principle. This suggests that it might be an improvement to give extra votes to the less well off. I suspect that this would lead to abuses, also. But I see no way to rule it out *a priori*. Even if it is true that there is no superior alternative to the now well-entrenched norm of one-person, one-vote, that is not something we could know without considering the results of relevant alternatives.

When democracy was first tried in the United States, it did not have much of a track record. The evils of hereditary monarchy were enough to motivate an experiment with a new form of government. We now have a much better idea of the advantages of a democracy, when democracy is understood broadly to include a package of civil and political rights. Amartya Sen's (1999) discovery that democracies do not have famines was one of the major developments. Kant [1795] correctly predicted a democratic peace: Democracies are much less likely than any other form of government to go to war with one another (Weart 1998). In addition, there is a "democracy advantage" in economic development (Halperin et al., 2005).

This *ex post* discovery of the advantages of democracy is an example of one of the reasons that democracy has advantages over other forms of government. Like economic markets, democracies can be thought of as political markets in which politicians market themselves and their policies to voters. Competition assures that the political parties will give the voters what they want. Burke [1790] thought that would be the downfall of democracy, because politicians would give voters what they wanted rather than what was good for them. It is easy for us to see in retrospect that Burke was mistaken, but it is not so easy to say why he was mistaken.

We may not have a full explanation of why he was mistaken, but it seems to me that the explanation has three important parts: (1) the free give-and-take of opinion, (2) the claim of first-person authority, and (3) the willingness of most people to incur at least modest costs to promote equity.

The Free Give-and-Take of Opinion

As I explained in chapter 7, one of the grounds of human rights is that human beings have no direct access to truth. Our best access is through a social process of free give-and-take of opinion. This is true in any area of inquiry, and so it gives democracies a knowledge advantage in any area of inquiry. But there is one kind of question for which this knowledge advantage is particularly important.

Governments endorsed by the main principle promote the life prospects of their citizens. They must do this even though we have no definition of

well-being or direct insight into what the best life for human beings is. Democracies can be thought of, in part, as ongoing social experiments to determine the answer to this question. For the process to make progress over time, no one has to start out knowing the answer to the question. What is necessary is the same kind of ability that is necessary for markets to promote life prospects, the ability to judge successes and failures in particular cases and to learn from them. When there is free give-and-take of opinion, over time, we all get to learn from the successful and unsuccessful experiments of others and they, too, can learn from ours.

This process of learning from positive and negative feedback is the same process that gives democratic governments an advantage over other forms of government. Any government that has the power to suppress negative information about its policies will use it. Only democracies effectively limit that power. Democracies work not because citizens can accurately predict the effect of government policies, but because the government cannot prevent them from finding out the information they need to evaluate whether the government's policies have worked and then using that information in deciding how to vote.

To understand the process of free give-and-take of opinion, it is helpful to keep in mind the model of markets. In a market it is not necessary that everyone know how to build a carburetor for everyone to benefit from that knowledge. But it is necessary that buyers be able to reliably evaluate cars. In part, they rely on experts to do this. *Consumer Reports* is trusted by so many buyers because it does independent expert testing of consumer products. But for it to be successful, it has to conduct tests on characteristics that consumers care about. *Consumer Reports* does not evaluate products on the basis of environmental footprint. If it had done so 20 years ago, it would have become a much smaller specialty publication. However, it is not at all implausible that sometime in the future, consumers may care enough about a product's environmental footprint that *Consumer Reports* will have to include that factor in its evaluations. In that way, *Consumer Reports* is like a democratic government that adjusts its environmental policies on the basis of feedback from its customers, the voters.

In a democracy, voters rely not only on experts, they also rely on the opinions of other voters. In a free market, businesses have always depended on word-of-mouth recommendations to generate sales. In the Internet era, this phenomenon has reached an entirely new dimension. In the past, people could obtain word-of-mouth recommendations only from acquaintances. In 2010 on Amazon.com, you can find evaluations of almost any consumer product you might want to buy from people you don't know. The result is a single number, an approval rating, based on one-person, one-vote. And so, the analogy between markets and democratic governments is even stronger today than it was a few years ago, because now there is a way that people's votes can provide feedback on the products they buy. This is just another example of the fact that markets and democracies are the same kind of solution to similar, though not identical, problems.

The Claim of First-Person Authority

All the knowledge in the world would not enable democracies to promote life prospects if people were not reliable judges of their own well-being—that is, if the claim of first-person authority were not true. Though far from infallible, it is people's judgments of how well their lives are going that provide the crucial feedback necessary to keep the government's policies sensitive to whether or not they are promoting life prospects. First-person authority is also the ground of a right against paternalism, so I discuss it more fully in chapters 12 and 13.

We should not confuse the claim of first-person authority with the claim that promoting one's own self-interest is a good life for human beings or with the claim that citizens in a democracy should vote according to self-interest. In a democracy, people will be free to live a life based on self-interest, but it would be unfortunate if very many people chose such a life.

And if everyone in a democracy chose to vote on the basis of self-interest, the results would be dire. In the late eighteenth century, the opponents of democracy predicted that it could not work, because it would devolve into a struggle for the power of a majority to tyrannize a minority. If voters simply voted their self-interest, democracy would be a license for a majority tyranny. So there is one more element that plays an important role in securing democracy's advantage.

The Willingness of Most People to Incur at Least Modest Costs to Promote Equity

Majority tyranny is more than an abstract possibility. It maintained a stable system of legally enforced segregation in the southern United States for decades. Legally enforced segregation ended because enough protesters, mostly black, were willing to pay large personal costs and a relatively impartial audience of citizens, mostly white, outside the South were willing to pay relatively small costs to promote fairness, in part by voting for candidates and policies that opposed legal segregation. In my first volume (Talbott, 2005, 148–157), I explained why I believe that a democracy could not be stable unless most people were willing to incur at least some small costs to promote fairness. It is also one of the motivational sources of the tendency of democracies over time to improve themselves as evaluated by the main principle. As I mentioned in chapter 9, it is also playing a prominent role in changing the incentives for businesses to make equitably promoting life prospects a contributor to their bottom line.

All three of these elements—the free give-and-take of opinion, the truth of the claim of first-person authority, and the willingness of most people to incur at least modest costs to promote equity—enable democratic governments to solve the reliable feedback and appropriate responsiveness problems. But no system is perfect. There is always room for improvement.

The Time Lag Problem

Early critics of democracy may have been mistaken in their predictions that in a democracy the majority in power would inevitably expropriate the wealth of the minority. But there is a similar kind of expropriation problem that even the initial opponents of democracy did not anticipate, expropriation from future generations.

Earlier in this chapter, I mentioned the U.S. government's budget deficits in current accounts, Social Security, and Medicare that have the potential to saddle our children with debts for the government services that their parents received before they were born as well as for their financial support and medical care in retirement. How did this happen? How did the lure of tax cuts blind so many voters to the consequence that their children—or, even worse, *other people's children*—would be paying the principal and interest on the loans that paid for those services and benefits? I think the answer is that there was no way to see or hear from the people who will be the victims of the inequity. If adults from the future could have been teleported into the present to make appearances on TV and radio talk shows and to ask questions at presidential debates, I believe that their voices would have been heard and most voters would have responded to them. The problem would be even easier to solve if the members of future generations had voting rights on policies, such as long-term debt, that directly affected their interests.

If the advantage of democracy is that it provides feedback on policies based on their effects, then one of the problems of democracy is that there is no way to get feedback on policies whose effects are far in the future. Of course, this is a problem for any form of government. Call it the *time lag problem*. The response to climate change is another example of this problem (e.g., Gardiner forthcoming). The problem of expropriating from future generations might be solved by a constitutional amendment requiring not only a balanced budget, but full funding of future liabilities. But there is no fix for the general problem. Until the technical problems of communicating with future people can be solved, the time lag problem will continue to be a serious problem for democracies.

When Tom Brokaw (1998) wrote a book about his and my parents' generation, he called it *The Greatest Generation*. Tom Brokaw and I are in the same generation. I shudder to think what title our children will give to the story of our generation. *Après Moi, Le Deluge*?

There still may be time to avoid the ignominy of being the first generation to have left our children (and their children, etc.) a lower standard of living than we have had. The time lag problem is a new kind of test for democratic governments. So far, they (and we) are earning a failing grade. But there is still hope. Throughout history far more commentators have underestimated the problem-solving potential of democracies than have overestimated it.

The Corrupting Effect of Private Campaign Financing

The time lag problem seems to be an issue for all democracies. One of the most serious problems for U.S. democracy, the corrupting effect of private campaign financing, is not. It is a local problem. Many democracies don't have the problem. For example, in Canada, campaigns receive public funding and individuals are allowed to make private contributions and then within strict limits.

In the United States, campaign contributions are sometimes thought of as a form of bribery. This is a mistake. Bribery involves promises of *quid pro quo*. For campaign contributions to be effective, no such promises are necessary. Large donations have an implied threat effect, because they signal to the candidate that if he does not act according to the donor's wishes, the donor could transfer the donation to a competing candidate. It is true that a system that operates by implied threats rather than by promises will be somewhat less effective and thus somewhat less corrupt, but these are matters of degree. The corruption is still endemic.

It is so obvious that this is a problem for U.S. democracy that it is worth considering why there have not been more dramatic changes. Here we confront a potential limitation of the gradual change favored by the main principle. Certain social practices have the functional effect of attractors. They do not actually attract anything, but the parties' motivations are such as to make it seem as though they have great powers of attraction. Campaign financing abuses are an attractor of this kind.

A system of private campaign financing will work fine if the financing is based on small contributions from individual donors. However, if larger contributions are permitted, then those who have more of a stake in certain legislation will be motivated to make larger contributions to those candidates who support that legislation. Notice that at this stage, there is no corrupt motivation. However, already the system is being corrupted to favor the interests of larger contributors over smaller contributors.

The individuals become corrupted when the larger contributions begin to operate as implied threats to fund a competing candidate if a legislator does not support desired legislation. Notice that it is perfectly reasonable that contributors would want to contribute to candidates who supported legislation they desire. Nonetheless, the process and the candidates themselves will ultimately be corrupted by that reasonable desire.

This corrupt process is an attractor because there is a strong tendency for any system of unregulated private campaign financing to evolve into it. It is a stable attractor, because once the system has become corrupt, the only hope of changing it is to somehow motivate the very legislators who have been rewarded by the system with lavish campaign contributions to vote to terminate those contributions. There have been some experiments with campaign finance reform in the United States, but they have all been less effective than originally anticipated. What is worse is that even though

protection of the value of one-person, one-vote should be one of the highest priorities of the U.S. Supreme Court, the Court has ruled many limits on private campaign expenditures to be unconstitutional limitations freedom of speech.[11] As I discussed in chapter 7, there should be no presumption that freedom of speech entails an unlimited right to amplify one's message (e.g., in advertising). The fact that the U.S. Supreme Court has interpreted the right to political speech to include a right of amplification has made the corrupt political system in the United States a paradigmatic stable attractor.

Legislators and the Main Principle

In chapter 9 I explained why the main principle does not endorse judges' applying it in their reasoning. The main principle is not a ground-level legal principle. What about legislators? Should they apply the main principle in making laws?

There are two ways of taking this question. First, should legislators be allowed to think about how to modify existing law to do a better job of equitably promoting life prospects? Of course. But no system of democratic rights would be endorsed by the main principle if the system depended on legislators' being motivated to equitably promote life prospects. Legislators will have a multiplicity of aims—most prominently, the aim of getting reelected—that will influence their positions on legislation. To be endorsed by the main principle, a democratic system must tend to produce laws that equitably promote life prospects even when the individual legislators have more mundane motives. For the system to work, it is essential that enough of their constituents be willing to incur small costs to promote equity. But most of the time, on laws that do not address important issues of equity, constituents will want their legislators to vote their interests and, in doing so, the legislators will generally adopt legislation that tends to enhance life prospects more than it reduces them.

Although there is no reason for legislators not to think about how to equitably promote life prospects, the main principle would not endorse giving their opinions on the subject more weight than the decisions of judges based on principles of constitutional adjudication. When there is a conflict between majority will and individual rights, it is almost irresistible for legislators to come down on the side of majority will. After all, if they don't, it is very likely that in the next election they will be running against someone who will. Democracies thus select for legislators who will weight the majority will over individual rights. Because individual rights are so important to the equitable promotion of life prospects, the main principle would support allowing legislators to make that determination only if there were no reasonable alternative.

Collective Action Problems and Human Rights

Most of the rights on my list of human rights are solutions to CAPs. A democratic government itself is a solution to a CAP. Once a democratic government is established, it can legislate solutions to lots more CAPs. Not all solutions to CAPs establish legal rights, but many of them do. Of all the potential legal rights that can be established as solutions to CAPs, which ones qualify as human rights? The test I have been employing is that the rights should be universal, in the sense that they are solutions to a CAP that every society confronts; they should be robust, in the sense that they should hold even against a majority and thus would typically be afforded constitutional protection; and they should be inalienable, in the sense that people should not be able to voluntarily relinquish or sell or trade them.

Democratic rights are part of a package of political rights that includes constitutional protections of the human rights, a democratic procedure for adopting legislation, and an independent judiciary to interpret and apply the constitution. By themselves, democratic rights are not sufficient to assure that a government will tend to equitably promote well-being. But when made part of a constitutionally protected package of human rights that includes the autonomy rights, it is the only form of government that can be relied on to do so. No one would have predicted this *a priori*. Indeed, there probably would not be any democracies if authoritarian forms of government had been equitable promoters of well-being, rather than misery.

Equity Rights

> We cannot be content, no matter how high [the] general standard of living may be, if some fraction of our people—whether it be one-third or one-fifth or one-tenth—is ill-fed, ill-clothed, ill-housed, and insecure.
>
> This Republic had its beginning, and grew to its present strength, under the protection of certain inalienable political rights—among them the right of free speech, free press, free worship, trial by jury, freedom from unreasonable searches and seizures. They were our rights to life and liberty.
>
> As our nation has grown in size and stature, however—as our industrial economy expanded—these political rights proved inadequate to assure us equality in the pursuit of happiness. We have come to a clear realization of the fact that true individual freedom cannot exist without economic security and independence.
>
> —Franklin D. Roosevelt

> A prosperous society could guarantee everyone medical care, education, decent housing, unemployment insurance, child care allowances, retirement benefits, and even a minimum income. It is entirely imaginable, in other words, that one might constitutionalize the elimination of poverty. . . .
>
> —Thomas Nagel

The main principle favors the *equitable* promotion of life prospects. What does equity require? Rawls argued that it required maximizing the position of the least well off. As I discussed in chapter 4, this requirement seems too extreme.

G. A. Cohen (2008) thinks it is too lenient. Cohen recommends changes in our moral beliefs to make people more egalitarian. Think of how different the world would be if all successful entrepreneurs donated the great preponderance of their wealth to promoting life prospects of the least well off, as Bill Gates and Warren Buffet have committed to do. The main principle would endorse such a change if it were done by persuasion. Because there is no prospect of this happening in the foreseeable future, it is worth considering what kinds of legal guarantees aimed at promoting equity—that is, *equity rights*—would be endorsed by the main principle.

Equity rights include two broad categories of human rights: opportunity rights and social insurance rights. *Opportunity rights* include rights to negative opportunity, nondiscrimination, and rights to positive opportunity, best understood in terms of the development of one's capabilities (e.g., Nussbaum 2000; Sen 1999). *Social insurance rights* include rights to health insurance, unemployment insurance, disability insurance, and maintenance (e.g., food stamps, a housing allowance, and some kind of welfare benefit).

Opportunity Rights

Life prospects are in part a function of the available opportunities and what one does with them. Negative opportunity is nondiscrimination. Positive opportunity is the capability to engage in gainful employment and the other activities of a normal life. The combination of negative and positive opportunity rights assures a person that hiring decisions will generally be based on relevant capabilities and that he or she will have the relevant capabilities for some jobs.

Rawls's requirement of *fair equality of opportunity* combines both negative and positive opportunity, in contrast to what he calls *careers open to talents* (1971, 72ff.). Fair equality of opportunity is surely a laudable ideal, but we can set it aside, because the main principle's endorsement depends on a substantive evaluation of a practice and on an evaluation of the practice of implementation. There is no way of implementing anything close to fair equality of opportunity, for reasons that I take up next. So if we want a name for the kind of opportunity rights endorsed by the main principle, it will have to be something less inspiring. I suggest this: *not too unfair inequality of opportunity*.

Opportunity rights provide a good example of how the main principle favors incremental improvements rather than radical transformations in the name of an ideal. Radical transformations are almost always disastrous.

It does not take great insight to recognize the greatest impediment to equality of opportunity. It is the fact that parents are motivated to invest in their own children and those investments can be very unequal. Plato recognized the problem and saw that there was a solution that would equalize investment: Remove children from their parents at birth and raise them in common.

If no one knew who their children were, no one could favor their own child. Would the main principle favor such a radical change? Surely not. First, to be endorsed by the main principle, it would have to be endorsed as a substantive practice. Think of what this would entail. It would require us to change a system of parental investment in children, which produces high levels of voluntary investment in children by parents, with a system that presumably would provide housing, care, and education to children for at least 18 years all financed by taxes, because the amount of voluntary investment would almost surely be negligible and it would require a huge government bureaucracy. It is hard to imagine a proposal that would more

dramatically reduce the life prospects of children. Even worse, it would significantly reduce the life prospects of their parents, at least of those who wanted to care for their children.

Second, the proposal must be evaluated as a practice of implementation. I have no idea how such a proposal could be implemented without causing a parental revolt. Would the army be called in to suppress the revolt? What would keep the army from revolting, since they are parents, too?

I conclude that on both types of evaluation, the proposal would be greatly disfavored by the main principle. This conclusion is reinforced by the fact that the main principle favors gradual improvements in the status quo over wholesale change, unless the changes are of the same kind as changes that have been successful elsewhere. Radical changes in social practices usually have unanticipated effects that can have disastrous consequences. The French Revolution and the Communist revolutions are historically the most important examples. But the evidence from experiments with "utopian" communities is also relevant. Such experiments almost never survive more than a generation, because it rarely takes more than a generation for the unanticipated disadvantages of the experiment to become evident, and even if the first generation is able to maintain its utopian zeal, the second generation typically does not. The main principle endorses a social practice of allowing small voluntary communities to conduct these experiments, but it would not endorse using coercion to impose a radical, untried experiment on an unwilling population.

Negative Opportunity Rights: Nondiscrimination

A negative opportunity right is a right not to be discriminated against on arbitrary grounds. I focus on discrimination in employment. Much the same can be said about discrimination in education, housing, public accommodations, or other areas in which discrimination can significantly reduce life prospects.

Antidiscrimination law illustrates again the importance of evaluating laws as policies. There are two different ways of formulating antidiscrimination laws. One way would be to prohibit all discrimination in employment except discrimination based on factors that are relevant to job performance. Although well-intentioned, such a law would be a nightmare to enforce. It would rule out hiring family members or friends or someone from your hometown or someone who graduated from the same college that you did or someone who has the same hobby you do or not hiring a redhead because your former spouse was a redhead. Even if you think that hiring decisions should not be based on such factors, it is easy to see that a law that permitted a legal challenge to any hiring decision thought to include such factors would invite huge numbers of lawsuits. The costs of enforcement would be very high.

Would the costs of enforcement be justified by the main principle? Only if the lack of such a law would significantly lower some people's life prospects.

But the sorts of discrimination mentioned so far are relatively harmless, because they do not significantly reduce people's life prospects. Not to be hired because of such a factor is upsetting, but it is not the kind of factor that would prevent a person from being hired in a comparable position.

From the point of view of policy, laws against discrimination justify the costs of enforcement only when the discrimination is of a kind that significantly reduces some people's life prospects. This leads to narrowly focused rights against discrimination, limited to those factors that, as a matter of fact, do or have significantly reduced people's life prospects. I refer to such factors as factors involved in *systematic discrimination*. In the United States, the factors of race, color, sex, national or ethnic origin, religion, age, and disability have been involved in systematic discrimination, so they are now routinely included in antidiscrimination laws. Sexual orientation has been added in some jurisdictions. This seems to raise a new issue, because discrimination based on sexual orientation is typically based on behavior, which is something that a person has control over. In this respect, sexual orientation is more like religion than the other items on the list of standard kinds of systematic discrimination.

Because religion and sexual orientation are both central areas of personal autonomy, it is easy to see that the main principle would endorse prohibitions on discrimination on either of those two grounds. Thus, the rationale for rights against discrimination on the basis of religion or sexual orientation is very much the same as the rationale for liberty rights against legal paternalism, which I discuss in chapter 13.

Are there other kinds of systematic discrimination that should be legally protected? In Australia, it is illegal to discriminate in employment on the basis of height, weight, or physical appearance (unless there is a reason to specify such requirements because of the nature of the job).[1] There is evidence that discrimination on all three of these factors significantly reduces earnings. On that basis, such discrimination would be categorized as systematic and protection against such discrimination would be endorsed by the main principle.

Positive Opportunity Rights: Capabilities

After the protection of security rights, there is no greater contribution to life prospects than guarantees that children are able to develop the capabilities that will enable them to take advantage of opportunities to find suitable work and to cooperate with others in mutually beneficial joint projects. The development-of-judgment rights are crucial, because they are the rights necessary to develop good judgment. They include the following:

1. A right to physical security
2. A right to physical subsistence

3. Children's rights to what is necessary for normal physical, cognitive, emotional, and behavioral development, including the development of empathic understanding
4. A right to an education, including a moral education aimed at further development and use of empathic understanding

Nussbaum has provided a list of 10 central human capabilities (2000, 78–80). Although not identical to the rights on my list, as applied to children, they are very much equivalent.[2]

Once most children have been assured what is necessary to develop their capabilities for good judgment, they will have everything they need to be able to become healthy, productive adults, so long as they receive the education and training to prepare them to enter the workforce.

How much education would be needed? That would depend on the educational requirements of the workforce. It would have to be enough to provide young people with a real choice of a career. From behind the veil of ignorance in the expanded original position (EOP), the life prospects of someone with no other choice than a career at a repetitive, robotic, minimum wage job would be so bleak as to make improving those life prospects a matter of real urgency.[3]

What about children with disabilities? When feasible, positive opportunity rights would entitle children with disabilities to special education and special accommodations necessary to enable them to develop the capabilities for good judgment and to be able to take advantage of opportunities for suitable work.

Although there are many elements to a good life, because the main principle evaluates life prospects on a narrow conception of well-being, it will evaluate alternatives on the basis of a narrow set of capabilities: capability for health, to participate in social life, to marry and have a family, and to engage in productive employment, with a reasonable choice of careers. In a capitalist economy, lifetime earnings are a good proxy for these capabilities, so comparisons of median lifetime earnings are a good way of evaluating social practices. This is true, because, as a general rule, bad health or inability to participate in social life will generally have an adverse impact on lifetime earnings and earning power makes it possible to marry and have a family. Of course, those with disabilities will still have special needs, covered by the other equity rights, but if those with a given disability were found to have the same median lifetime earnings as those without it, that would be a good indication that they had developed the capabilities covered by the positive opportunity rights.

Capability rights are endorsed by the main principle, because investing in the development of children's capabilities is one of the best investments a society can make for promoting life prospects. How does the main principle evaluate the status quo, in which parents with very different resources invest very unequal amounts in developing the capabilities of their children? For all

its inequality, because the status quo benefits from parents' willingness to voluntarily contribute to the development of their own children, most children are assured of the resources to develop the necessary capabilities.

What about the children who are not? Providing the resources necessary for children of deprived backgrounds to develop their capabilities is an urgent demand of equity. Where would the money come from to fund the necessary programs? An obvious source would be inheritance taxes. Because children of wealthy parents already have greater opportunities than they would have if they had been born to poor parents, from the point of view of equity, it seems perverse that, in addition, they should receive large inheritances whereas children of poor parents receive little or no inheritance. How high should the inheritance tax be? It seems to me that the main principle would favor setting it at whatever level would maximize inheritance taxes. There is no reason that the main principle would require it to be lower, and making it higher would be counterproductive.

How is the United States doing in providing not too unfair inequality of opportunity? A rough indicator is the U.S. Census Bureau reports on median household income. Here is the report for 2007 by racial/ethnic category:

1. Asian: $66,103
2. White, not Hispanic: $54,920
3. Hispanic $38,679
4. Black $33,916 (DeNavas-Walt et al. 2008)

These numbers are not strictly comparable because, for example, the average household size of Asian families is higher than the average household size of white families. Also, many Hispanics are first generation immigrants and the second generation can be expected to have higher earnings. It is the figures for black household income that are most disturbing. They indicate that, in 2007, the United States was still a long way from not too unfair inequality of opportunity.

Social Insurance Rights

In a modern economy, social insurance rights include rights to affordable health care, to disability insurance, to unemployment insurance (discussed in chapter 9), to retirement insurance (i.e., to what in the United States is called *Social Security*), to some sort of maintenance allowance for those whose incomes are very low (e.g., food stamps, subsidized housing) or for those who have no other source of income (i.e., what in the United States is called *welfare*). Social insurance rights are often regarded as the newest and most controversial category of rights, but in fact they are the oldest and least controversial. Human communities exist to provide social insurance. Every traditional society has some system of social insurance. Most traditional

societies guarantee all members of the community a social minimum. In a poor community, this minimum may be secured by loans that must be paid back. Or it may be secured by community action, as, for example, in farming communities in the United States where, if a neighbor's barn burns down, the entire community turns out to rebuild it.

One of the reasons that doubts have been raised about human rights to social insurance is that the level of such insurance necessarily depends on the level of wealth of the society. However, this is no bar to identifying social insurance as a category of rights, because the main principles standard of equity will necessarily require higher levels of insurance in wealthier societies, to keep the inequalities from becoming too large.

My discussion will address social insurance rights in a modern economy, because implementing the economic and political rights on my list would generate such an economy. I focus my discussion on disability insurance, because it raises many of the most important theoretical questions. I discuss the other kinds of insurance more briefly.

Establishing a Social Floor for Life Prospects: Disability Insurance

The most important kind of insurance for persons with disabilities would be provision for special education and special accommodations to enable persons with disabilities to be productive members of society. Some disabilities are too severe to make this a practicable goal. For these disabilities, insurance would be in the form of an annuity, for those who did not have other means of supporting themselves.

As discussed in chapter 4, Rawls's theory does not apply to these people, but only to those who are "normal and fully cooperating members of society over a complete life" (1993, 20). It was plausible for him to limit the scope of his theory this way, because the theory was intended to articulate a standard of moral reciprocity, and it seemed to him that those with severe disabilities would not be able to reciprocate cooperation. However, this seems to me to be a mistake. The main principle also articulates a principle of moral reciprocity, but its scope can include those with severe disabilities if we think of them as members of the class of nonresponsible noncompliers. The default assumption is that they would have been willing to cooperate had they not been disabled, so they should be included within the scope of moral reciprocity. On this basis, they are included in the EOP.

Once they are included, it becomes an easy matter to show that the main principle would endorse disability insurance. Consider the question from behind the veil of ignorance in the EOP, as described in chapter 4. The main principle requires that life prospects be evaluated at every stage of life. At birth, life prospects for those with severe disabilities would be extremely low. Rawls had no way of incorporating them into his theory, because if they

were included in his theory, it would be clear that an index of primary goods was not a remotely reasonable proxy for life prospects and that maximin was not the appropriate principle for determining the level of insurance. But in the EOP, though insurance would be provided for those with severe disabilities, even if they were the least advantaged group, there would be no reason to maximize their life prospects, if this would drastically reduce the life prospects of others. The main principle gives some priority to the life prospects of the least well off, but not absolute priority.

There is one more thing to notice about disability insurance. The level of disability insurance would clearly depend on the general level of wealth in the society. Only in a very wealthy society, much wealthier than any society that exists today, would the level of disability insurance equal the amount of damages for disability that would be awarded in a legal action for negligence or strict product liability. Recall the discussion of the example of paralysis caused by a defective product in chapter 4. I said there that the main principle could endorse different levels of compensation for paralysis from different causes. We can now understand why.

Consider the difference between someone with a congenital disability for which no one is responsible and someone with the same disability due to injury by a defective product or some other liability in tort. The former would receive disability insurance at a level determined by the EOP test. The latter would receive disability through a products liability court action in tort law, in which the legal standard would be full compensation. This level of full compensation will almost inevitably be substantially higher than the level of disability insurance, because only in societies much wealthier than any existing society would the EOP test set the level of disability insurance equal to the legal definition of full compensation.

Why is the legal standard for liability for defective products full compensation? The answer depends on understanding how the system of strict liability for defective products functions to make the economy a self-regulating system, as I explained in chapter 9. In the economic system, the costs of compensating those injured by defective products are spread over all those who purchase the product, because manufacturers must raise their prices to cover these costs. Full compensation is necessary so that the external costs of the product can be internalized in its price. If a price that internalizes these external costs is too high and purchasers are unwilling to pay it, then the product will not be able to "pay its own way" and it will be eliminated.

Comparison to Dworkin's Hypothetical Insurance Market

It is useful to compare this account with R. Dworkin's (2000, 76–81). Dworkin also argues for disability insurance, but he reaches that conclusion by a different route. The main difference is that Dworkin sets the amount of insurance not on the basis of an original position test, but on the basis of an empirical,

though counterfactual, question about what level of disability insurance the parties would actually have purchased at market rates, if they didn't know whether they had the disability. When the question is whether to buy insurance for a severe, congenital disability, the question makes no sense, because a person with this degree of disability would never reach the level of cognitive functioning necessary to make such a choice. So Dworkin makes the decision on the basis of the level of disability insurance that the average member of the community would have bought to avoid a comparable disability (2000, 78).

To see why the main principle would differ from Dworkin's proposal, let us focus on a 21-year-old who has already been provided with the level of disability insurance that the average member of the community would have bought. Under Dworkin's proposal, the 21-year-old would be free to increase his insurance, decrease it, or leave it as it is. We can assume that some 21-year-olds would choose each of the three alternatives, because the policy represented the *average* amount of insurance that someone would purchase, and we would expect some individuals to prefer more and others to prefer less than the average. Indeed, Dworkin's proposal leaves it open that someone might cash in the policy and use the funds for something else. Imagine, for example, that Robert, a 21-year-old, has good reason to believe that his parents would support him if he ever became disabled, so he cashes in his disability policy and all other insurance policies that he owns to use as part of his start-up funding for a new business. The new business is promising, but just as he gets it going, there is a recession, and he loses everything. Shortly after, while driving in a car with his parents, the three are involved in a collision that is his father's fault. His father and mother are both killed, and he is severely disabled. In a subsequent negligence action, his parents' estate is awarded to the occupants of the other car, who were severely disabled in the accident, also. So Robert is disabled, destitute, with no source of income and no entitlement to any social insurance. On Dworkin's account, this would be an example of a loss due to option luck, so Robert would have no claim of justice, and thus no legal claim to support of any kind.[4]

Aware of such a possibility, Dworkin opens the door to the possibility of the state acting paternalistically to prevent Robert from selling his insurance policies in the first place (2000, 217–218). This move actually threatens Dworkin's entire construction, because the entire compensation system depends on granting moral authority to the choices that people make and would make. But I can allow that there might be some cases in which decisions to go uninsured could be reversed on paternalistic grounds. I just don't see how this is one of them. Paternalistic intervention must be based on some irrationality in the target's choice. It seems to me that we can fill out the example so that Robert's choice is quite rational. After all, it was probable that his business would succeed and that he would eventually have plenty of money to support himself if he became disabled and it was very probable that his parents would have been able to support him if his business failed and he later became disabled. To say that the government was justified in intervening

in Robert's decision on paternalistic grounds would imply that the government is justified in intervening in all of our lives to prevent us from engaging in any activity that would increase our chances of a very bad outcome, which includes just about everything we do. So it is hard to see how the government could be justified in preventing Robert from cashing in his insurance policies on paternalistic grounds.

Now this would seem to raise a problem for my view, because my view is that rights to disability insurance should be inalienable—that is, I also hold that Robert should not be permitted to cash in his government-provided insurance for disability. I can't explain the inalienability of this right in the way that I explained the inalienability of democratic rights and other human rights in chapter 10. Robert is not in a collective action problem; prohibiting him from cashing in his disability insurance is not a solution to a collective action problem. So am I not in the same boat as Dworkin? Don't I have to find a way to justify paternalistic intervention to prevent him from cashing in his disability insurance?

The answer is no, because I did not define the level of disability insurance as a function of what Robert would be willing to buy. The main principle evaluates distributions of life prospects on the basis of a standard of equity, not on the basis of what gambles a person would make. To see that there is a difference, imagine an alternate scenario in which Robert's package of disability and other insurance had been paid for not by the government, but by his Uncle Sam. Uncle Sam bought the policies when Robert was born. They are lifelong policies. When Uncle Sam bought the policies for Robert, he specified policies that could not be redeemed for cash or assigned to a third party. Now, at age 21, Robert is trying to raise investment capital. His uncle has no extra capital that he can invest in the business. They have the following conversation:

ROBERT I don't understand how you could have treated me so paternalistically as to buy me insurance policies that I can't cash in. Don't you realize that I have a very promising business opportunity that needs start-up funds. If I ever do become severely disabled, my parents will be able to support me.

UNCLE SAM My decision to buy policies that could not be cashed in was not paternalistic. My goal never was to maximize your life prospects. That is your goal, and you are doing a good job of it. I am not questioning your judgment. If I were you, I would want to be able to cash in the policies, also. My goal in purchasing the insurance was to protect you against the possibility of very improbable but very bad outcomes. If I had bought you insurance policies that you could cash in, you would have been able to increase your life prospects by cashing them in, but I would have failed to achieve my goal, which was to provide a level of protection against something very bad happening. I was protecting you against bad *outcomes*, not maximizing your life *prospects*.

I think Uncle Sam's explanation of why his purchase was not paternalistic matches the rationale for social insurance endorsed by the main principle. The main principle endorses social insurance to establish a floor below which no one can fall. If the social insurance policies are alienable, some people will cash them in. Even if everyone who cashes them in is rational in so doing, some of them will be unlucky and, as a result, will fall through the floor. Making the policies inalienable is the only way to establish a *social floor without holes*.[5] If this is the goal of social insurance, then making social insurance policies inalienable is not paternalistic.[6] Uncle Sam's rationale is the government's rationale under the main principle for making social insurance policies inalienable.

This distinguishes my account from Dworkin's, but it seems to raise a parallel problem. Recall that I argued that if Dworkin permitted government paternalism to prevent Robert from cashing in his policies, then the government would be permitted to prevent almost any action that involves some risk of a very bad outcome. So the question for me is this: If the main principle endorses limiting Robert's freedom to cash in his insurance policies, why doesn't it also endorse limiting the liberty of anyone who acts in a way that increases the risk of a very bad outcome?

To answer that question, let's return again to the EOP. It is clear that limiting people's liberty to take actions that increase the risk of a very bad outcome would have devastating effects on life prospects. So that policy would not be endorsed by the main principle. But there is another important difference as well. As a general rule, *limits* on individual autonomy are not favored by the main principle. But the main principle does not require that governments *maximize* individual autonomy. This example shows us that there is a different kind of role that governments can play, to use social insurance to establish a social floor without holes. If social insurance policies are aimed at establishing a social floor without holes, it is no criticism of them that they do not maximize autonomy. So when we view the alternatives from the point of view of the EOP, we see that equity can favor a social floor without holes without favoring other limits on autonomy.

There is another way that Dworkin's account, which makes each individual's level of social insurance a function of her attitude toward risk, is at odds with the main principle. The main principle might favor a society that encourages risk taking by entrepreneurs. Of course, most new businesses fail, but the few that survive generate great benefits for the society. If the society promotes risk taking, in Dworkin's scheme, as a side effect it would reduce the level of social insurance that people would choose at the same time that it increased the need for it, because risk takers would generally purchase a lower level of insurance than the risk averse, but many of them would be in the position of needing it when their businesses failed. In contrast, the main principle would endorse setting a level of insurance independent of a person's attitude toward risk and then making it inalienable.

Dworkin's objection to such proposals is that they permit the society to spend money on things that the members of society would think are irrational (2002, 124). This objection is an artifact of his own theory, which makes social choice a function of individual choices. When we tease apart social choice from individual choice, we can see that there can be two different standards of choice that one person can consistently hold. The social choice is not aimed at overruling individuals' own judgments about how to live their lives (which would be paternalistic); it simply has a different goal from the individual's choice, the goal of establishing what I have called a *social floor without holes*. It is quite consistent for one and the same individual to favor individual choices in his own life that risk falling through the holes, but to favor a social policy that provides a floor without holes.

The final difference between Dworkin's account and my account was discussed in chapter 9. On Dworkin's account, once everyone has reached the equality of resources baseline, there would be no justification for redistribution of income from the more well off to the less well off that would increase the life prospects of those just above the baseline, no matter how great the inequalities were. Thus, although Dworkin does not concur with Nozick's general analysis of the Wilt Chamberlain example, he does concur with Nozick's conclusion that there is no justification for any additional redistribution from Wilt to others, so long as everyone is above the baseline (2000, 111). The main principle does not take the level at which people would choose to self-insure to have any moral significance. Because of the priority given to the life prospects of the less well off, as a general rule, it regards as an improvement in equity a practice (e.g., a progressive income tax) that can transfer wealth from the more well off to the less well off without altering their motivation to engage in productive activity.[7]

Health Insurance

Any modern economy without affordable universal health insurance can easily be improved under the main principle. No such society could satisfy the most minimal standard of equity without guaranteeing affordable protection against serious health threats. Because the main principle gives priority to the less well off, it will favor practices that prevent or cure disabling and fatal conditions. Because it evaluates *life* prospects, other things being equal, it will tend to favor preventive care for children and young adults over care that temporarily extends a declining old age.[8] If children could vote, there would have been much more adequate and universal health insurance for children in the United States long ago. This is another example of how democracy does a much better job of representing the life prospects of those who can vote than of those who cannot. In any case, mandatory affordable health insurance is an essential part of providing a social floor without holes.

Retirement Insurance

The main principle would not require the government to fund retirement pensions, but it would require some mandatory system of funding of retirement pensions. Funding of retirement pensions is a worldwide problem for democracies in the developed world, because of the time lag problem that I discussed in chapter 10. So far as I know, no country in the world follows the accounting standards for reporting future liabilities and for funding them that are routinely required for private businesses. In the United States, the government and the press typically misreport the government current accounts deficit, by adding in the tax receipts intended to fund Medicare and Social Security, without any reference to the liabilities that those receipts are supposed to fund. This is not just a failure of the government and the press. Politicians may exploit the citizen animus against taxes, but it is ultimately the citizens' willingness to incur debts that future generations will have to pay that is the problem. No form of government, not even democracy, is very good at responding to problems in the distant future. Perhaps the only solution is to adopt a constitutional amendment requiring a balanced current accounts budget and full funding of future liabilities.

There is no requirement that retirement insurance be in the form of a government pension. One advantage to private retirement accounts is that there is no way for the government to spend or borrow against them. But if they are to provide a social floor without holes, then individual owners would also have to be prevented from spending or borrowing against them, except perhaps in extreme cases. Also, investments would have to be restricted to a small number of relatively safe investments.

Would such restrictions be paternalistic? Is it paternalistic to make retirement deductions mandatory rather than optional? As we have already seen, the answer is this: not necessarily. If the goal is to provide a social floor without holes, the rationale would not be paternalistic. However, I do think there is some paternalism behind these proposals. So the question is whether the paternalism is justified. That is a question that I won't be able to fully answer until the chapter 13, but the main idea is simple. If, on reflection, people either approve or come to approve of the government's treating them paternalistically, then the paternalism does not overrule their judgment about what is good for them; it gives effect to it.

Rights to Maintenance: Food Stamps, Subsidized Housing, and Welfare

I discuss rights to maintenance in two parts. First, I discuss support such as food stamps and subsidized housing for low-wage earners. These programs fit under the redistribution that I discussed in chapter 9. They alter the distribution of income in a market economy that is based on commodity

value and make it more equitable. So long as such programs are administered efficiently, they would obtain general agreement in the EOP and are easily endorsed by the main principle.

Rights to welfare are more controversial. There would be less need for welfare in a society that did a better job than the United States does of developing the capabilities of its children for productive lives and gainful employment. But I think it is clear that the EOP would favor some kind of temporary assistance for young adults to further their education or to obtain training to improve their employment prospects. Also, I think it is clear that the EOP would provide assistance during periods when those who were mentally ill were undergoing treatment. Those with severe and intractable mental illness would be covered by disability insurance.

The difficult question is whether there should be a system of insurance that pays for the maintenance of able-bodied and -minded adults. Many countries have much more generous rights of this kind than the United States. Would the main principle endorse such rights? It would obviously not endorse a welfare benefit that was so generous that it motivated a significant number of workers to quit work and live on welfare. So any system of welfare rights has to be sensitive to the insurance effect, the effect that a system of insurance makes the contingency insured against less undesirable, and thus potentially motivates people to act in ways that generate more of it. But that is an issue about the level of benefits. A more fundamental question is whether an able-bodied and -minded citizen should have a right of this kind at all. There is in the literature an argument for a right of this kind, a right to an *unconditional basic income* that anyone would qualify for if they had no other income. How would the main principle evaluate this proposal?

Right to an Unconditional Basic Income?

Van Parijs (1995) has proposed that an unconditional basic income be guaranteed to everyone. Every adult, rich and poor, employed and unemployed, would receive a regular income of some amount to be determined. This is not an unprecedented idea. Alaska uses its oil revenues to provide a resource dividend to every adult resident of Alaska. In 2008 the amount of the dividend was $3,269. This is a substantial amount, more than the maximum federal earned income credit.

It should be clear that the main principle would evaluate such a proposal on the basis of its consequences. It might very well have good consequences, especially in promoting equity. By providing the benefit to everyone, van Parijs would eliminate almost all of the costs of administering a welfare system, because most of those costs are incurred in administering the hurdles that applicants are required to jump over to make sure that they qualify. In a basic income system, everyone would qualify.

However, there are two worrisome aspects of the proposal. The first is illustrated by Van Parijs's defense of it and by G. A. Cohen's defense of a similar position. This is the problem of *victimization*. The second problem is the problem of *benefit spreading*.

The problem of victimization is not necessarily a problem with the basic income proposal itself, but it is a problem with a certain kind of argument that has been used to defend it. Call it the *argument from inequality*: Do you have trouble getting yourself out of bed in the morning? Have no motivation to do anything other than hang out on the beach and surf? Have trouble finding someone who will marry you? Indeed, are you unhappy with any of your genetic endowments? If so, you are a victim. You are entitled to compensation. The unconditional basic income is a way of correcting an inequity. It is not charity. You are entitled to it. Society owes you.

How should we think about this argument? It is true that we are not responsible for our genetic endowment. It is also true that some people have it easier in life because of their genetic endowment. If there is a God, perhaps this could be the basis for a claim to compensation against him, a claim for genetic disadvantage. To make it the basis for a claim in our world would be the end of the idea of moral reciprocity. Moral reciprocity is the idea that those who participate in a cooperative endeavor are entitled to a fair share of the benefits of cooperation. But this argument establishes entitlements with no contributions. According to this argument, society owes you. Why? For existing? I agree that if you are willing to contribute to society in some way, then it owes you a fair share of the benefits. How could it owe you for existing?

And if it does, there really is no need to drag yourself out of bed or to give up your surfing to find a job or to try to make yourself into a person someone would want to marry or to figure out how to best use your genetic endowments, because you are a victim. For as long as human societies have existed, they have depended on motivating their members to make contributions for the common good and have rewarded them for those contributions, not always equitably. Now we are told that the next stage of moral development will be one in which society owes us, just for existing! If this is a moral improvement, the main principle is no principle of moral improvement.

Though it is hard to see how the victimization argument could be correct, this is not necessarily fatal to the basic income proposal. It would be fatal to the basic income proposal if the effect of the basic income were to take away people's motivation to engage in productive activity. Van Parijs thinks there is no danger of that, if we start with a modest basic income. I think he may be right about this. And I can even see how a basic income could help to reinforce relations of moral reciprocity. It is a well-established human response that when given a gift, we are inclined to reciprocate. This was the motive behind the Hare Krishna followers who used to give out flowers to travelers at airports. Many people found it difficult not to reciprocate this unsolicited

gift. Providing everyone with a basic income might well generate feelings of this kind and make people more willing to make contributions for the general good. In that case, its motivational effects would be endorsed by the main principle.

It is also possible that it would free people to engage in artistic and other poorly remunerated activities that can produce great social benefits. However, I am not confident that it would have these motivational effects. That is one reason that I favor a negative income tax as a transfer from higher to lower wage earners, rather than a transfer that includes everyone. Low-wage earners are doing their part in the scheme of social cooperation.

There is another problem with the basic income proposal, the problem of benefit spreading. This problem can be explained by comparison to cost spreading. When the consumers of a dangerous product pay a high enough price to compensate those who are injured by the product for their injuries, the monetary cost of the injuries is shifted from the injured and is *spread over* all the consumers. Benefit spreading is the reverse. It occurs when benefits that are targeted at those who suffer a loss are reallocated and *spread over* the larger population of those who suffer a loss and those who don't (Talbott 1988).

Let me explain why I think that the basic income would have this benefit spreading effect. I will say that a government benefit that is provided only to those who suffer a certain kind of loss is a *targeted* benefit. The basic income would not be a targeted benefit. It would go to all adults. What would be the effect of instituting such a proposal? The cost would depend on the generosity of the benefit. Van Parijs advocates that we start small and then increase the benefit gradually. However, to provide a benefit even of the size of the Alaska resource dividend ($3,600) to all adults in the United States would cost nearly a trillion dollars. Van Parijs points out that this benefit would substitute for a lot of targeted benefits that the government now provides. That is my worry. As the basic income benefit grew, like a sponge it would soak up funding from all other targeted benefit programs. The effect would be to replace relatively larger benefits, targeted at those who have suffered a loss or disadvantage, with a relatively smaller benefit spread over the entire adult population. This would have the effect of shifting benefits from the more worse off to the less worse off. Because the main principle gives priority to the life prospects of the more worse off, other things being equal, it would not endorse such a shift. For this reason, though I could imagine a world in which the unconditional basic income proposal would be endorsed by the main principle, I don't think the main principle would endorse it in our world.

One more thing. The Alaska resource dividend is like manna from heaven, so it would be covered by allocative justice and an equal distribution to everyone at least comports with one conception of allocative justice. Income redistribution is an issue of distributive justice. It is not like manna from heaven.

Welfare

So I believe that any welfare proposal should be a means-tested proposal and should require that recipients pursue education or job training or be looking for employment. The most difficult question for any such system is what to do about mothers with young children. Were it not for the insurance effect, the answer would be easy. Mothers of young children should have some means-tested source of support for their children when their children are young. However, because of the insurance effect, such a welfare benefit will increase the number of single mothers who have children they can't support. This is a difficult issue. A threshold test for the adequacy of any attempt to address it should be the acknowledgment that it is a difficult issue. Authors who think that the answer is obvious—either because it is obvious that benefits should be provided to the child or because it is obvious that women shouldn't be rewarded for having babies they can't support—by their failure to see the difficulty of the problem invalidate whichever answer they give to it. I don't have an answer. I think we need to experiment with novel approaches.

The Inalienability of Equity Rights

Someone contemplating selling her vote is in a recognizable CAP. If other voters don't sell their votes, a single vote seller might be better off if she sells. However, if everyone sells, the result could be very bad.

Individuals contemplating the sale of their equity rights could be in a CAP, but at least for the social insurance rights, they need not be. The example of Robert, the entrepreneur who wanted to be able to cash in his social insurance for investments in a new start-up, illustrates why, even though the situation is not a CAP, the main principle still endorses making the social insurance rights inalienable and thus preventing Robert from making what might be a quite rational investment decision. The grounds are not paternalistic, because they do not depend on thinking that Robert's decision is irrational. The grounds are considerations of equity: to provide a social floor without holes.

TWELVE

The Most Reliable Judgment Standard for Soft Legal Paternalism

This chapter has a different structure from chapters 6 to 11. In those chapters, I was able to briefly explain the consequentialist rationale for the rights in question and then proceed to compare the consequentialist and nonconsequentialist accounts. In this chapter, I have the problem that there really is no adequate consequentialist account of a right against legal paternalism, especially an account of where to draw the line between soft (permissible) and hard (impermissible) legal paternalism. So I use this chapter to explain the consequentialist rationale for a new ground-level principle of soft (permissible) legal paternalism, the most reliable judgment standard, and compare it to the most influential nonconsequentialist account, that of Feinberg (1986). Then in the next chapter I use the most reliable judgment standard to explain what kind of liberty rights against legal paternalism would be endorsed as human rights by the main principle.

A Consequentialist Case for Liberty Rights against Legal Paternalism

What is legal paternalism? This is not an easy question to answer. Rather than delve into the complexities, I am going to use a rough-and-ready definition that will sort the cases of interest in the correct way. As I use the term, intervention in a target population's action is *paternalistic* just in case it is intended to promote the good of the target population by *overruling* their own judgment about what is good for them.[1] *Legal paternalism* is the enactment and enforcement of paternalistic laws. Generally speaking, when a paternalistic law is enacted, there will be a target population (those whose judgments are being overruled) who will regard themselves as worse off than they would be without the law and practically no one will regard themselves as significantly better off. The reason is simple. Consider a prohibition on going to movies on Sunday. Those who think it is better for them not to go to movies on Sunday can refrain from going whether or not there is a law. So the law does not make them any better off.[2] Those who would go to movies on Sunday but for the law are the target population. They will regard themselves as worse off with the law than without it.

It is important to distinguish legal paternalism from legal solutions to collective action problems (CAPs). Paternalistic laws resemble legal solutions to CAPs in that both sorts of laws are aimed at promoting the good of those they coerce. However, although legal paternalism *overrules* the target population's judgments about what is good for them, a legal solution to a CAP *gives effect* to the target population's judgments about what is good for them by bringing about an overall outcome they generally regard as better for them than the outcome that would eventuate if there were no law. For example, if there were no legally enforced traffic signals, almost all drivers would regard themselves as worse off, because driving would be so much more hazardous.

In evaluating a policy of legal paternalism, it will be important to focus on the attitudes of the *target population*—that is, those whose judgments of their own good are overruled by the law—at the time of the *intervention* in their actions—that is, at the time in which it forces them to do something that they would otherwise not do or prevents them from doing something that they otherwise would do.

The nonconsequentialist has an easy explanation of why there should be a right against legal paternalism. Paternalistic intervention is incompatible with respecting autonomy.

Because there is no adequate consequentialist account of the conditions for justified legal paternalism in the literature, this chapter is devoted to developing one. The challenge is to develop an account that will not merely yield the same results as the much simpler nonconsequentialist account, but will do a better job of drawing a line between soft (permissible) and hard (impermissible) legal paternalism. In this chapter I consider several different proposals for where to draw the line between soft and hard legal paternalism.

The nonconsequentialist bases the case for a right against legal paternalism on autonomy. What does the consequentialist case depend on? The answer is that it depends on the reliability of people's judgments about their own well-being. At first glance, this seems like a flimsy foundation for a right against paternalism. What paternalist ever proposed intervening in people's decisions when their own judgment about what was good for them was reliable? Paternalists seek to intervene in decisions when people's judgment about what is good for them is *not* reliable.

Mill was the first consequentialist to try to base the case for a right against legal paternalism on the claim that, given appropriate background conditions (specified by the other human rights), normal adults generally are reliable judges of what is good for them—at least, more reliable than the government. This is what I call the *claim of first-person authority*.[3] So far as I know, Mill was the first person to make this claim. But there is a puzzle about his argument for a right against paternalism. To appreciate the puzzle, consider a typical assertion from Mill's argument: "But the strongest of all the arguments against the interference of the public with purely personal conduct is

that when it does interfere, the odds are that it interferes wrongly and in the wrong place" ([1859], 94).

The puzzle is that it does not follow from the fact (presuming that it is a fact) that the "odds are" that the government interferes wrongly, that *every* piece of paternalistic legislation does so (cf. G. Dworkin 1972). It seems that Mill's conclusion, that the government should never intervene paternalistically, is a non sequitur. But it is not. To resolve the puzzle, it is necessary to realize that Mill was an *indirect* utilitarian. An indirect utilitarian evaluates the practice or policy of enacting paternalistic legislation on the basis of the expected utility of the *policy*. Probabilistic information of the kind that Mill cites is exactly what is needed to evaluate a policy of this kind. For Mill's argument to go through, he needs to show that the expected utility of a policy of legal paternalism is lower than the expected utility of an antipaternalistic policy. This is all that Mill could reasonably claim, because no utilitarian could claim that paternalistic intervention is *never* successful. But if the *policy* of legislating paternalism has worse consequences than the antipaternalistic policy (and if there is no subclass policy of legislating paternalistically for which the inequality is reversed), then Mill can consistently hold both that some paternalistic laws might well increase overall utility, but, because governments can't know in advance which laws will do so, the best policy for maximizing overall utility is the antipaternalist one. Of course, it is open to Mill's opponent to reject one or more of his empirical claims.

At the time that Mill made his argument, many people thought it was obvious that the claim of first-person authority was false. Mill's claim is not generally accepted today. However, as I mentioned in chapter 9, economic markets would collapse if people's judgments about what is good for them weren't fairly reliable.

In any case, if the claim of first-person authority is true, then it can be the foundation of a consequentialist case for a right against legal paternalism. So I am going to proceed on the assumption that it is true and then evaluate that assumption more fully in chapter 14.

If the claim of first-person authority were true, it might seem that it would rule out all legal paternalism, but that is not right. It will, however, provide the foundation for a distinction between soft and hard legal paternalism that allows very little of it to qualify as soft when directed toward normal adults who would not consent to it. Because the consequentialist distinction between soft and hard paternalism is not the same as the distinction drawn by nonconsequentialist accounts, it provides us with another test of the two kinds of account. However, although nonconsequentialist accounts of the distinction tend to be fairly simple, the consequentialist account is more complex. It will take me the rest of this chapter to work it out and to critically compare it with nonconsequentialist accounts. Once that has been done, I can take up the question of what sorts of rights against legal paternalism it would support in the next chapter.

Almost everyone agrees that some paternalism is justified—for example, paternalism toward young children and toward the severely mentally ill or severely mentally impaired. Some of those who think that some paternalism can be justified nonetheless oppose all *legal* paternalism, either because of the potential for abuse (e.g., the involuntary commitment of political opponents as mentally ill) or because of the unavoidable side effects of legally enforcing paternalistic laws (e.g., the large number of those currently incarcerated in the United States for selling or using small amounts of illegal drugs, where the laws are aimed at protecting drug users whether they want to be protected or not).

These concerns with potential abuse and potential negative side effects are very real practical concerns that would be relevant to any attempt to justify legal paternalism. However, I wish to temporarily set these practical concerns aside, so I can focus on what might be called the *pure theory of legal paternalism*, that is, the question of what sorts of paternalism could be justified if there were no concerns about potential abuse or potential negative side effects. Because my account carries no presumption against legal paternalism for children and nonautonomous adults, in the remainder of this chapter I focus primarily on paternalism targeted at adults with normal development of normal cognitive and emotional capacities, who are autonomous in the nonmetaphysical sense that I discuss more fully below. The best way to explain my consequentialist account is to contrast it with Feinberg's (1986) influential nonconsequentialist account. I begin, however, with a stricter standard.

The Explicit Voluntary Endorsement Standard for Soft Paternalism

Some would hold that paternalism is never justified, unless the target of the paternalistic interference has previously given her explicit, voluntary consent to it. Call this the *explicit voluntary endorsement standard for soft paternalism*. It is surely right that explicit voluntary consent can justify paternalistic interference, but most people would not believe it reasonable to define justifiable paternalism so narrowly. Suppose, for example, in a feverish state, Arnold hallucinates a pedestrian bridge outside his third-story window, where there is none. Suppose he decides to walk right out the window. Fortunately, you and some of his other friends are visiting. When you realize what he is planning to do, you try to talk him out of it by pointing out the consequences of falling from a third-story window. Arnold replies by telling you that you have overlooked the fact that he will be walking on a sturdy bridge that can easily support his weight. Nothing you say can convince him there is no bridge. You and his other friends know that if you forcibly prevent him from walking out the window, Arnold will be grateful to you after his fever passes. So when Arnold tries to step through the

window, you and his other friends physically restrain him. Then you call the legal authorities, who confine Arnold for his own protection for 24 hours until the fever resolves.

Undoubtedly, you and Arnold's other friends and the authorities who were called all interfered with Arnold's liberty for his own good. Because he disagreed at the time, your intervention required you to overrule his own concurrent judgment about what was good for him. Was the intervention justified? According to the explicit voluntary endorsement standard, it would be justified only if Arnold had previously given his explicit voluntary consent to it, perhaps by having executed a durable power of attorney giving you the authority to make decisions for him during periods of incapacity. This is unrealistic. Most people have not given their explicit consent to all the various kinds of paternalistic interference they would endorse if they were asked. This suggests a modification to the explicit voluntary endorsement standard.

Feinberg's Voluntariness Standard for Soft Paternalism

The preceding example suggests that paternalism can be justified by hypothetical or implicit consent as well as by actual, explicit consent. This is the idea behind what is surely the most influential account of soft legal paternalism in the literature, Feinberg's (1986) voluntariness standard.

Feinberg's account of soft legal paternalism focuses on the choice the paternalistic intervention aims to prevent. On Feinberg's account, "the state has the right to prevent self-regarding harmful conduct when but only when it is substantially nonvoluntary, or when temporary intervention is necessary to establish whether it is voluntary or not" (1986, 126). Feinberg has no formula for when conduct is "substantially nonvoluntary." He allows that the threshold varies with the level of risk, the irrevocability of the risked harm, and other contextual factors (1986, 118–122). The test that he finds most useful and the one that is relevant to the examples that I discuss is that the agent's conduct "represents him faithfully in an important way, expressing his settled values and preferences" (1986, 113). Applied to the example of Arnold, Feinberg's account would hold that Arnold's choice to walk out the window was substantially nonvoluntary because it did not express his settled values and preferences. Intervention to prevent him from walking out the window would qualify as soft paternalism because it would in fact promote Arnold's own settled values and preferences, which were to live, not to die.

Feinberg applies his account not only to cases involving temporary incapacity or derangement, but also to cases of simple ignorance. Thus, he agrees with Mill on the *example of the unsafe bridge*:

> If either a public officer [Dick] or anyone else saw a person [Harry] attempting to cross a bridge which had been ascertained to be unsafe,

> and there were no time to warn him of his danger, they might seize him and turn him back, without any real infringement of his liberty. . . . [4]

In the example of the unsafe bridge, Dick is assuming that Harry does not intend to put his life at risk. Dick could be mistaken. Suppose that after Dick stops Harry from crossing the bridge, when Dick explains to Harry that the bridge is unsafe, Harry replies he knows it is unsafe. He wants it to collapse because he is a movie stuntman making a film. On Feinberg's account (as on Mill's) intervention would no longer be justified and Dick should allow Harry to proceed (1986, 125).

There is one further complicating factor in Feinberg's account that I must mention, because it closes a loophole that might otherwise vitiate the account. Suppose that after Dick prevents Harry from crossing the bridge, Harry decides that it is important enough to him to get to the other side that it is worth the risk of crossing the bridge. We could imagine Dick replying as follows: I understand your values better than you do. In my judgment, crossing the bridge would not express your settled values and preferences (and therefore would not be voluntary); therefore I am authorized to prevent you from crossing the bridge. It is clear that if the government were permitted to intervene whenever it thought that your actions did not express your settled values and preferences, this would open the door to a large amount of legal paternalism.

The simplest solution to this problem would simply be to specify that voluntary conduct is conduct that faithfully expresses the agent's settled values and preferences *in accordance with what the agent's own stable judgment of how best to further them (in light of the available information) would be*. Feinberg does not adopt this simple formula because he is willing to allow expert knowledge to overrule the agent's own judgment (1986, 131). Because the examples I discuss do not involve agents who refuse to accept expert opinion, they are cases in which Feinberg would agree with the results of this simple formula (133–134). So I employ the simple formula in my discussion.

The Hypothetical Endorsement of Intervention Standard

There is one improvement that I propose to make to Feinberg's account.[5] It concerns what exactly the test for soft legal paternalism is to be applied to. In a potential case of legal paternalism, Feinberg's test focuses on the choice to be interfered with. I believe it would improve his test to focus it on the target's attitude toward the intervention.

I illustrate the difference with a simple example. Consider suicide. Suppose that most suicides are substantially nonvoluntary, because they are the product of a temporary period of depression. I do not believe that this settles the question of whether intervention to prevent suicide is strongly paternalistic.

I believe that it depends on the agent's attitude toward the relevant kind of intervention. Suppose that most people when they are not depressed would voluntarily endorse a law that permitted medical personnel to order a temporary period of involuntary commitment when they are diagnosed with suicidal depression. The fact that they would voluntarily endorse involuntary commitment would make the intervention qualify as soft paternalism even if they would not endorse the involuntary commitment at the time of the commitment proceedings.

The reason that their voluntary endorsement should be part of the standard is illustrated by continuing the example. Suppose that it is possible to screen people's medical records and reliably identify those who are at risk of becoming suicidally depressed. If there were electronic medical records, the government could screen everyone's medical records and identify those who are at risk. Suppose there were a way to involuntarily medicate those who were at risk by dissolving antidepressants in the water supply of those who were at risk. And suppose that if the government were to do so, the suicide rate would go down by 10%. Although it is easy to imagine that a large majority of the population would voluntarily consent to a suitably limited policy of involuntary commitment, it is hard to imagine that a majority would ever voluntarily consent to a policy of involuntary medication. This leads me to conclude that the test for soft/hard paternalism should focus not on the target's conduct but on the target's attitude toward the relevant policy of intervention. Thus, I propose the following amendment to Feinberg's voluntariness standard:

Hypothetical Endorsement of Intervention Standard for Soft Legal Paternalism. Legal paternalism is soft whenever the target would *voluntarily* endorse the policy of intervention (i.e., whenever, in light of the available information, the target's stable judgment about how to further his settled values and preferences would favor the relevant policy of intervention).

The hypothetical endorsement of intervention standard explains why, in the example discussed above, involuntary commitment would be soft legal paternalism, but involuntary medication would be hard legal paternalism. Even if it is an improvement on Feinberg's standard, is it an adequate standard of soft legal paternalism? No, it is not. The reason is that, like Feinberg's standard, it defines the difference between soft and hard legal paternalism solely in terms of the target's actual or hypothetical beliefs, values, and attitudes *before and/or at the time* of the paternalistic intervention. I refer to such standards as *backward-looking standards*. I use the hypothetical endorsement of intervention standard as a proxy for all backward-looking standards, including Feinberg's.[6]

As a condition for soft legal paternalism, the hypothetical endorsement of intervention standard is neither necessary nor sufficient. It is not necessary, because some legal paternalism is soft even though the target would not hypothetically endorse it at the time of intervention; it is not sufficient,

because some legal paternalism that the target would hypothetically endorse at the time of intervention would not be soft. The most important exceptions to the standard are the exceptions to necessity—that is, cases in which legal paternalism is soft, even though the target would not hypothetically endorse it. Most of my discussion will focus on examples of that kind. I take up sufficiency later.

The Hypothetical Endorsement of Intervention Standard Is Not a Necessary Condition for Legal Paternalism to Be Soft

In order to develop a consequentialist standard of soft legal paternalism, I am going to introduce a useful expository tool that will enable me to explain some of the main ideas. The tool will not enable me to precisely state the consequentialist standard, only to approximate it. So I need to distinguish between the preliminary statements of the standard, which will employ my expository tool, from the more precise final statement of the standard, which will dispense with the expository tool. The expository tool will use branching diagrams to represent hypothetical lives, as I explain shortly.

We are looking for a consequentialist standard for soft legal paternalism. Consider the Arnold example again. An obvious candidate for the relevant difference between the judgment of the feverish Arnold that he can walk out of the window and the judgment of the nonfeverish Arnold that he cannot is that the Arnold's nonfeverish judgment about his own good is more reliable than his feverish judgment. So perhaps what is significant about voluntary choices is that they are reliable. This would provide an alternative explanation of the reason for outsiders to favor Arnold's nonfeverish judgment over his feverish one. This would also connect the discussion of paternalism with the discussion of the consequentialist value of choice in chapter 9.

I want to pursue the idea that the significance of voluntary choices is their reliability with some additional examples. To keep the examples simple, I assume that the relevant choices are those of adults who are autonomous, in the consequentialist sense that I explain below, and that the agents' endorsements are based on their judgments of what is best for them. I am going to see what kind of standard of soft legal paternalism we are led to if we think of voluntariness in terms of the reliability of one's judgments about one's own good.

My first example will be hypothetical, but I want to ground it in a real-world example. Suppose there is a recreational drug RD that does not directly cause harm to anyone other than those who take it. Allen is an autonomous 21-year-old. He welcomes new experiences and wants to live life to the fullest. Allen is aware of studies showing RD to be dangerously addictive, but he knows lots of people his age who use it and who strongly recommend it. They seem to have suffered no ill effects. So he judges that, on balance, it would be good for him to take RD, also.

To make the example interesting, I need to suppose not only that there are scientific studies on the effects of RD, but also that there are statistical studies on the attitudes of users and former users toward the drug. We do have some information of this kind for cigarette smoking. One survey reports 82% of those who smoke believe it would be better for them if they did not and most of them have tried many times to stop, without success.[7] This figure is almost surely a low estimate of the percentage of those who currently smoke who will later regret their decision, both because it includes many smokers who have not been smoking long enough to come to regret it and because it does not include all the former smokers who have successfully quit. It is reasonable to suppose that a very large percentage of those who do successfully quit would wish they had never started. So I think it is not unreasonable to suppose that 90% or more of those who become cigarette smokers someday will regret ever having started.

The reason I must move from this real-world example to a hypothetical one is that it is not reasonable to suppose that such a high percentage of those who take up smoking would eventually come to endorse a legal prohibition on smoking, even if the potential for abuse and the negative side effects of the prohibition were minimal. Most people think it is better for them to be free to make mistakes and to learn from them than to be prevented from making any. However, some mistakes are so tragic that most people would want to be prevented from making them. Suppose the recreational drug RD has such devastating effects on most people's lives that within 20 years of beginning to use it, 90% of users will not only regret making the decision to use the drug, they will also judge it would have been better for them had there an effective prohibition on using it. In such a case, I will say that they will come to *endorse* the prohibition. Assume, also, the evidence shows that their judgments endorsing a legal prohibition are quite stable over time. Once they come to endorse a legal prohibition, they don't change their minds.

To make the example complete, I must consider what the attitudes of potential drug users would be if there were a prohibition. Suppose that, although they would be prevented from using the drug, they would still be able to obtain information about the effects of the drug and that 90% or more of them would come to endorse the existence of the prohibition. These are strong assumptions. Shortly I show how to relax them.

Consider again the case of Allen, a 21-year-old who plans to take the drug. There is no drug prohibition, so Allen can use the drug if he chooses to. Allen has heard about tragic cases involving the drug, but all his friends who use it seem glad to be taking it. He knows that over 90% of those who use the drug come to regret the decision within 20 years, but he attributes that change to their growing older. He is like the young man described by Nagel (1970, 74), who in his youth values sex, spontaneity, frequent risks, and strong emotions, but who expects that in 20 years he will value security, status, wealth, and tranquility.[8] He does not now endorse the values he expects himself to

have in 20 years, so it is no surprise that he does not now endorse the judgments he expects himself to make in 20 years either.

What are we to say about this case? Would it be permissible for the government to intervene to prevent Allen from using the drug, even though his decision to use it is based on his stable judgment, after reviewing all of the available evidence on the drug, about how best to further his settled values and preferences? On the assumption that intervention would be effective in preventing use of the drug and there was no potential for abuses of the law and no other potential negative side effects of the law, it seems to me that a legal prohibition could be justified. Before I explain why, note that it could *not* be justified on the hypothetical endorsement of intervention standard. In this case, choosing to use the drug is in accord with Allen's stable judgment, based on all the available evidence, about how to further his settled values and preferences. Intervention to stop him from using it would thus be hard paternalism. However, I believe that the prohibition could be justified, at least in theory, not on the basis of *other people's* judgments that taking the drug would be bad for Allen and that intervention would be justified, but on the basis of what is reasonable to believe that Allen's *own future judgment* would be.

Given the statistical data postulated above, it would be overwhelmingly probable that in the future Allen would come to endorse such a law in two hypothetical cases: first, if there were no legal prohibition and he was not prevented from using the drug; second, if there were a legal prohibition and he was prevented from using it. It is useful to illustrate Allen's situation with a diagram (see figure 12.1).

The diagram's branches represent two different scenarios: One branch reflects his future if there is no paternalistic intervention (-PI) and Allen starts taking the drug; the other branch reflects his future if there is paternalistic intervention (PI) and he is prevented from taking the drug. Before he reaches the division, he does not endorse paternalistic intervention (-E). No matter whether there is paternalistic intervention or not, for some time

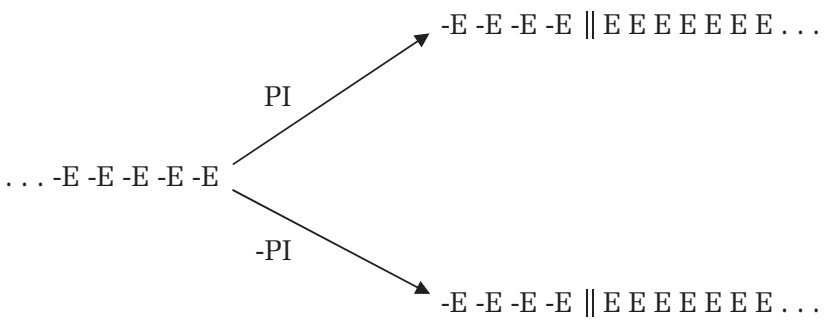


Figure 12.1. Allen's Decision Whether or Not to Use the Drug RD.

afterward he will continue to oppose such intervention (-E). In either case, however, eventually he will change his mind and come to endorse (E) paternalistic intervention. If there is no paternalistic intervention, he will start taking the drug and eventually come to the conclusion that it would have been better for him if he had been prevented from taking it; and if there is paternalistic intervention, he will not take the drug and he will eventually come to the conclusion that it was good for him to have been prevented from taking it. Finally, in either case, his change of mind to endorse paternalistic intervention will be *unequivocal*—that is, he will not later change his mind. Once he comes to endorse paternalistic intervention, he will continue to do so.

Of course, there is no way to be absolutely certain what Allen's future judgments will be. In the example, I assumed there was a sound statistical basis for being able to predict his future judgments with a probability of 0.9. Shortly I explain why any probability above 0.5 will typically be sufficient to justify legal paternalism.

On first impression, it may seem quite counterintuitive for me to claim to be able to justify paternalistic intervention to prevent Allen from taking the drug. There is such a strong presumption against paternalistic interference with a person's voluntary choices that it will take me some time to explain why intervention in Allen's case should be permitted. To do so, it is useful to talk about the judgments of Allen's current and future hypothetical selves. This talk of temporal selves is simply a useful heuristic. I do not mean to imply that Allen really is a series of different selves. After employing the heuristic, I explain how to dispense with it.

It is important to recognize that the door I am opening for exceptions to the hypothetical endorsement of intervention standard is a narrow one. I am suggesting only that, in this case, Allen's *own* actual or hypothetical future judgment can justify overruling his current judgment about what is good for him, even if his current judgment is based on his settled values and preferences. I am not suggesting *other people's* judgments about what is good for him could justify overruling Allen's own judgment.

The Most Reliable Judgment Standard (First Statement)

To justify legal paternalism in Allen's case, I have to give the judgment of Allen's future hypothetical selves priority over the judgment of his current self. How can this be justified? To answer that question, recall the consequentialist rationale for a right against legal paternalism. On the consequentialist rationale, rights against legal paternalism are justified by the claim of first-person authority, that, given the appropriate background conditions defined by autonomy rights, a normal adult is a more reliable judge of what is good for her than other people are. The claim of first-person authority leaves it open whether some of a person's own judgments about what is good for her

are more reliable than others. And thus it leaves open the possibility of a narrow category of justified legal paternalism that involves overruling a person's *own less reliable judgment* about what is good for her, in order to give effect to *her own more reliable judgment* about what is good for her. This suggests a new standard for soft legal paternalism. Here is an initial statement of the standard:

Most Reliable Judgment Standard for Soft Legal Paternalism—First Statement. Legal paternalism that intervenes in the action of a target T is soft if and only if it is reasonable to believe that, based on T's most reliable judgments of what is good for him, T endorses (or would endorse) intervention of the relevant kind.

How could the most reliable judgment standard support paternalistic intervention to prevent Allen from using the drug? Although there is no guarantee that our future selves' judgments about what is good for us are more reliable than our past selves' judgments, for most of our lives, our judgments about what is good for us become more reliable with time. This is true because human beings learn from experience.

It is clear that people typically regard their hypothetical future selves' judgments about what is good for them as more reliable than their current judgments. I say *typically*, because we all realize there is often a significant decline in memory and other cognitive functions late in life. Let me set aside such qualifications by focusing on future judgments made by a future self before any significant decline in cognitive functioning occurs. It is hard to deny some sort of priority to the judgments of such a future self. For example, think of how useful it would be in choosing a career to be able to consult with hypothetical future selves to find out what each career would really be like.

However, I must distinguish two ways that what we learn from experience might play a role in future endorsement. Consider Ed, who each week plays the lottery, in which the odds of winning are 1/1,000,000. Today, as Ed considers whether to buy a particular ticket with number *N,* he realizes that it is almost certain the number *N* will not be chosen and that his ticket will be a loser. Thus, he realizes that it is almost certain that his future self will regret his choosing number *N.* Should this information affect his decision today? As I have described the case, the answer is no. Of course, Ed's future self knows the winning number, so Ed's future self would like for Ed to choose the winning number. But Ed's future self does not oppose the practice or policy of playing the lottery. We can suppose that his future self still plays every week. This is not the kind of learning from experience that plays a role in the most reliable judgment standard.

The case would be different if Ed's future self had come to the conclusion that playing the lottery itself is a bad gamble and that it is better for him to spend his money on other things. Or if Ed's future self had read more about lottery winners and had come to the conclusion that winning the lottery is a

curse rather than a blessing and, thus, not something that he should aspire to. Or if Ed's future self had changed his values so that wealth was no longer a value for him (cf. Millgram 1997). In each of these cases, Ed would not simply be regretting the results of his *particular* choice, which is a common human condition. Instead, he would be finding reasons that would undermine his earlier reasons for thinking that engaging in the *practice* or *policy* of lottery playing was good for him, of which his particular choice is only one instance. These are the kinds of reasons that are relevant to the most reliable judgment standard.

It is important to recognize that even if Ed's future self came to regard the practice of playing the lottery as not good for him, it is unlikely that he would come to endorse intervention to prevent his earlier self from engaging in that practice. So the requirements of the most reliable judgment standard are quite strict. Applied to the example of Allen and the drug RD, they would require not only that Allen's future self think it bad for him that he took RD on a particular occasion (perhaps when he had a bad reaction to the drug) or that Allen's future self come to believe that intervention to prevent him from taking RD on that one occasion would have been good for him and not only that Allen's future self come to believe that it would have been better for him if he had never used RD at all, but that Allen's future self come to believe that it would have been good for him if he had been legally prevented from using the drug RD (or any other drugs with relevantly similar effects) at all.

The most reliable judgment standard coincides with the hypothetical endorsement of intervention standard in most cases, because in most cases a person's current stable judgment about how to best further his settled values and preferences (in light of the available information) is the best evidence for what her future judgment will be. So, for example, both standards endorse intervention to prevent feverish Arnold from walking out his third-story window. The most reliable judgment standard yields this result because feverish Arnold's judgment is not stable and it is reasonable to believe that when his fever passes, Arnold's stable (and more reliable) judgments will unequivocally endorse intervention of this kind as being good for him, which is to say that not only will he come to endorse intervention of this kind, but he will also not change his mind later and withdraw his endorsement.

The example of Allen and the recreational drug RD does present a conflict between earlier and later judgments. In this sort of case, the hypothetical endorsement of intervention standard and the most reliable judgment standard may give different verdicts on the permissibility of legal paternalism. Whether they do or not will depend on the details of the policy the drug prohibition represents. It would be hard to justify a drug prohibition that included draconian punishments for possession of small amounts of the drug. However, a drug prohibition requiring that users be given treatment rather than prison sentences might well be justified by the most reliable judgment

standard, though it would not be justified by the hypothetical endorsement of intervention standard.

Many people are reluctant to accept that interference with Allen's decision can be justified, due to fears about how such powers could be abused. I believe this is a legitimate concern, and one that would have to be considered in determining whether such a policy should be adopted, all things considered. Here I continue to set aside concerns about potential abuse and potential bad side effects, so I can focus on the most important considerations of pure theory.

One way of defending the hypothetical endorsement of intervention standard would be to insist that in a case like Allen's, intervention *is* objectionably paternalistic. If Allen himself does not regard his future judgment as more reliable than his past and present judgment, the defender of the hypothetical endorsement of intervention standard might argue, it would be objectionably paternalistic for the law to overrule Allen's own current judgment of the reliability of his future judgment and impose *its own* judgment of the reliability of Allen's future judgment on Allen. However, this is a potentially misleading description of the basis for intervention. Just as it is reasonable to believe that Allen's future self will judge that preventing him from using the drug RD would have been good for him, it is reasonable to believe that Allen's future self will also judge his earlier judgment opposing the prohibition to be less reliable than his own later judgment. There is a symmetry to Allen's earlier and later judgments that is illustrated in figure 12.2. For every judgment made by Allen's earlier self endorsing the judgment of his earlier self, there is a corresponding judgment made by Allen's later self endorsing the judgment of his later self.

Allen's Earlier Self	*Allen's Later Self*
(1a) judges that a prohibition on drug RD would be bad for him;	(1b) judges that a prohibition on drug RD would be good for him;
(2a) judges that his earlier self's judgment (1a) is more reliable than his later self's judgment (1b);	(2b) judges that his later self's judgment (1b) is more reliable than his earlier self's judgment (1a);
(3a) judges that his earlier self's judgment (2a) is more reliable than his later self's judgment (2b);	(3b) judges that his later self's judgment (2b) is more reliable than his earlier self's judgment (2a);
. . .	. . .

Figure 12.2. The Symmetry of the Disagreement between Allen's Earlier and Later Selves.

Resolving Conflicts between Earlier and Later Selves

Allen's earlier self disagrees with his later self on the benefits of using the drug RD and on which self's judgment is more reliable. The most reliable judgment standard sides with Allen's later self. The advocate of the hypothetical endorsement of intervention standard can claim to be neutral between Allen's earlier and later selves.

Is such neutrality appropriate? Consider again Nagel's (1970) example of the young man, call him Tom, who in his youth values sex, spontaneity, frequent risks, and strong emotions, but who expects in 20 years that he will value security, status, wealth, and tranquility. Is the appropriate attitude here not to take sides in the disagreement between Tom's earlier and later selves?

There is a consequentialist case for favoring the later self. Though Tom's earlier self does not understand how it could be reasonable for him to come to value security, status, wealth, and tranquility, Tom's later self could easily understand how it could have been reasonable for his earlier self to value sex, spontaneity, frequent risks, and strong emotions. Indeed, it is quite plausible to think Tom's later self would not endorse intervention to prevent his earlier self from acting on those values, because he would think it was important that his later change of values be based on his own judgment in response to his experience, not on the forcible intervention of others.

In addition, if neutrality is the proper attitude toward the disagreement between Tom's earlier and later selves, should we also be neutral if Tom's *earlier* self endorses a policy of intervention to prevent his *later* self from acting on his later self's values. If it were possible, his earlier self might precommit to a life of frequent risks, so as to prevent his later self from being able to avoid them. Should such precommitment strategies be legally enforceable? This is closely related to the issue raised by Parfit's example of the Russian nobleman (1984, 327). In Parfit's example, an idealistic young nobleman wishes to be able to precommit to distributing his inheritance to the peasants, because he believes that by the time he receives the inheritance, his ideals may have faded and he may decide to enjoy the wealth rather than redistribute it.

The Parfit example is not directly relevant to the current discussion, because it involves a judgment about which course of action is better for everyone, rather than a judgment about what is better for the nobleman himself. To turn the example into one concerning judgments about one's own good, suppose the young nobleman believes great wealth would be bad *for him*, but he is concerned that when he receives the inheritance, he will be so blinded by the self-serving reasons for enjoying it that he will not be able to see why it is bad for him to do so. For this reason, he wishes to make an enforceable vow to turn over his inheritance to a humanitarian organization.

The most reliable judgment standard can help to explain why such vows should not be enforceable. After trying out poverty for a while, the nobleman

will generally be in a better position to judge how good it is for him. So his later self will be in a better position to judge whether it would be good for him to give up the inheritance. This is not to say that an earlier self should never be able to make commitments binding on a later self. There is no problem about commitments of a kind endorsed by both the earlier and later selves. However, when the commitment is based on a judgment of one's own good and there is good reason to think that a later self would *not* endorse such a commitment, that can provide a good reason for not permitting the earlier self to enter into such commitments. As I discuss in the next chapter, this is the main idea needed to explain why voluntary slavery contracts, religious vows, and various other precommitment devices should not be legally enforceable, though ordinary contracts should be.[9]

We seem to have discovered an area in which the judgments of the later self actually do limit the choices of the earlier self, for vows of this kind are not legally enforceable. The same kind of reasoning applies to the question of whether Tom should be able to precommit his future self to a life of frequent risks. Such a precommitment would not be legally enforceable. We now have an explanation, based on the most reliable judgment standard, of why it should not be. The explanation depends on our thinking that the judgment of Tom's later self about what is good for him is more reliable than the judgment of his earlier self.

Let me refer to the claim that people's judgments about what is good for them generally become more reliable over time as the *time-relative version of the claim of first-person authority*. There are two ways of understanding why this claim would be true. First, it might be held to be true simply because people's factual beliefs become more reliable over time. Call this the *weak* version of the claim. Practically no one would deny this claim.

The *strong* version of the claim holds that it is true because *both* people's factual beliefs *and* their evaluative beliefs about what is good for them become more reliable over time. The most reliable judgment standard depends on the strong version.

I can't mount a full defense of the strong version here. Suffice it to say, that it depends on thinking that the model of learning from experience applies to both our factual beliefs and our evaluative beliefs about what is good for us.[10]

Let me emphasize that the most reliable judgment standard only gives priority to the future self's evaluative beliefs about one's own good. Where decisions involve considerations of what is good for other people, I don't assume that the same case can be made for the greater reliability of the judgments of the later self. That would depend on whether such capacities as the capacity for empathy improve or decline with age. It may well be that young people's idealism gives them a greater capacity for empathy.

If the strong time-relative version of the claim of first-person authority is true, it is a contingent truth. We could imagine beings who start life knowing everything there is to know about what is good for them and knowing

everything they need to know to be able to make choices that would best promote their good. Over time, their cognitive capacities decline, so that their later judgments about what goals they should pursue and how they should pursue them are typically less reliable than their earlier judgments. For beings of this kind, the most reliable judgment standard would favor their earlier judgments over their later judgments. Even for beings of this kind, the most reliable judgment standard would lead to counterexamples to the necessity of the hypothetical endorsement of intervention standard, because it would justify overriding a subject's *current* judgment opposing intervention in order to give effect to an *earlier* one endorsing it.[11]

The strong time-relative version of the claim of first-person authority explains why the hypothetical autonomous endorsement standard is inadequate. It is backward-looking. No backward-looking standard for soft legal paternalism can be adequate. An adequate standard must be forward-looking.

It is important to emphasize that when we ask whether a future self would endorse paternalistic intervention, we are asking about the future self's attitude toward a certain *kind* of paternalistic intervention in a certain kind of action. This was the moral of the example of Ed the lottery player, discussed above.

It is also important to emphasize that when we ask whether or not the future self would endorse the relevant policy of intervention, we are asking for the future self's judgment on whether or not the policy of intervention would be good for her. Usually, there is no significant distinction between *endorsing* the policy of intervention and *believing that it is good for oneself*, but sometimes there is. An example is in my discussion of forced medical care in the next chapter.

Future Bilateral Endorsement

Paternalists often attempt to justify their intervention by saying "Someday you'll thank me for this." Call this the *future gratitude condition*.[12] It is important to note that the most reliable judgment standard's requirements for overriding a person's own hypothetical judgment are more stringent than this. The future gratitude condition says nothing about what the target's attitude toward intervention would be if no intervention were to take place. Thus, it leaves open the sort of possibility represented in figure 12.3.

Consider the case of a proposed ban on pornography. Suppose that if there is no ban on pornography, most pornography users will never endorse a ban. Suppose also that if an effective ban on pornography is enacted, most people who would have been users without the ban will become prudes and will come to unequivocally endorse the ban. This is the situation illustrated in figure 12.3. The ban would satisfy the future gratitude condition, because it is reasonable to believe that if it were instituted, the target population would eventually come to unequivocally endorse it (represented by the top branch

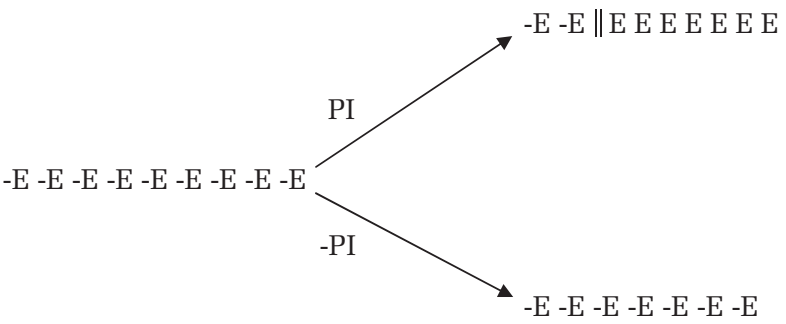


Figure 12.3. An example in which the future gratitude condition is satisfied but the paternalistic intervention is not weak, because of the lack of bilateral future endorsement.

in figure 12.3). But the most reliable judgment standard requires that, in a case such as this, both branches of the diagram lead to unequivocal endorsement of the ban.[13] The ban would not satisfy the most reliable judgment standard, because potential users who are not subject to the ban will never come to endorse it (represented by the bottom branch in figure 12.3).

Why does the most reliable judgment standard require both branches of the diagram to lead to unequivocal endorsement in order to justify overriding the earlier self's opposition to paternalistic interference? Because, in general, there is no principled basis for holding that the judgments in one branch of the diagram are more reliable than the judgments in the other branch. Each branch represents a future with different life experiences. There is no basis for claiming that one branch would produce more reliable judgments than those in the other branch.[14] So the most reliable judgment standard typically does not come into play unless it is reasonable to believe that both branches lead to unequivocal endorsement of paternalistic intervention. When both branches lead to unequivocal endorsement, I say the case is one of *future bilateral endorsement*. Future bilateral endorsement is represented graphically by the fact that both branches eventually lead to unequivocal endorsement of the relevant kind of paternalistic intervention. It is illustrated in figure 12.1.

Talk of earlier and later selves is a colorful way of talking about the attitudes that people would generally have in two different scenarios. The most reliable judgment standard would never diverge from the hypothetical endorsement of intervention standard unless it were possible to predict at least some future voluntary changes in attitude. To do so, in the ideal case, we would need to rely on relevant statistics. In the example of Allen, information about the trajectory of the judgments of others who have used the drug and those who have not provides evidence from which we can reasonably infer Allen's hypothetical future attitudes toward intervention to prevent him from taking the drug.

If it were 100% certain that potential users such as Allen would come to unequivocally endorse (UE) paternalistic intervention (PI), regardless of whether the intervention took place, that would be the strongest possible case for classifying the ban as soft legal paternalism. It would be unrealistic to think that 100% of a target population would ever endorse any kind of legal paternalism. How probable must it be that potential users of RD such as Allen would come to unequivocally endorse the intervention in each of the two branches, for it to qualify as soft legal paternalism? To answer that question, there are two probabilities that must be considered—first, the probability that a member of the target population of potential users would come to unequivocally endorse intervention, if the intervention were to take place (Prob(UE/PI)), and second, the probability that a member of the target population of potential users would come to unequivocally endorse intervention, if the intervention were not to take place (Prob(UE/-PI)).[15] What is the threshold value these probabilities must exceed for the intervention to qualify as soft paternalism?

Consequentialist considerations can help to answer that question. Consider this question: Which selection of a threshold value for the relevant probabilities would be the best policy of legal paternalism for equitably promoting the life prospects of the target population (in this case, potential users)? When the question is asked in this way, it can be seen to be the same kind of question as the one about what democratic decision rule would do the best job of equitably promoting life prospects. In chapter 10, I suggested that, when constrained by constitutional rights to protect minorities from majority tyranny, majority rule is probably the best policy for adopting most other legislation. It seems to me that similar considerations support the conclusion that the threshold value for most soft paternalism should be 0.5—that is, a policy of intervention toward a target population will typically qualify as soft legal paternalism on the most reliable judgment standard when Prob(UE/PI) and Prob(UE/-PI) are both greater than 0.5.[16] I refer to this as the *future bilateral majority endorsement standard*. I distinguish it from a different standard with which it might be confused in a note.[17]

When Prob(UE/PI) and Prob(UE/-PI) for the target population are both greater than 0.5, it is reasonable to suppose that more in the target population will benefit from the paternalism than will not. On average then, one would expect the gains from such a policy to outweigh the losses. Then it is reasonable to believe that the *policy* of enacting paternalist policies when Prob(UE/PI) and Prob(UE/-PI) are both greater than 0.5 would promote almost everyone's life prospects. It is important to emphasize that it is the *policy* of setting the threshold value at 0.5 that would be expected to promote *everyone's* (or almost everyone's) life prospects. In any particular application of the policy, some people's life prospects will be increased and some people's will be decreased. The *policy* of setting the threshold at 0.5 can improve everyone's life prospects, even if it is unrealistic to suppose any particular application of it would do so.

The analogy with majority rule yields another important insight about the future bilateral majority endorsement standard for soft paternalism. Suppose a policy of legal paternalism is enacted on the grounds that, for the target population, Prob(UE/PI) and Prob(UE/-PI) are both equal to 0.6. Then we can expect that after the policy is enacted, 60% of the target population will come to unequivocally endorse it and 40% of the target population will not. To the 40% who do not come to endorse it, the policy might seem to be hard legal paternalism, but it is not. It would be hard legal paternalism if the justification for the law depended on the 60% majority overruling the judgment of the 40% about what was good for the 40%. In the case as I have described it, the policy is justified by the future judgment of the 60% about what is good for the 60%. The adverse effects on the other 40% are not the goal of the policy, they are simply an unfortunate *majority spillover effect*. I give other examples of the majority spillover effect below.

Unilateral Endorsement

I have identified a category of cases that are classified as hard legal paternalism by the nonconsequentialist hypothetical endorsement of intervention standard, but qualify as soft legal paternalism by my consequentialist standard: cases that satisfy the future bilateral majority endorsement condition. These are cases in which satisfying the nonconsequentialist standard is not *necessary* for legal paternalism to be soft. Are there other exceptions to necessity? Yes, there is one other kind of case. These are cases in which the future bilateral majority endorsement condition is not satisfied, because the–PI branch terminates prematurely. See figure 12.4 for a diagram illustrating such a case.

Figure 12.4 represents a situation in which, if there is paternalistic intervention, the target will come to unequivocally endorse it, but if there is not, the target will lose the ability to endorse anything, which is represented by

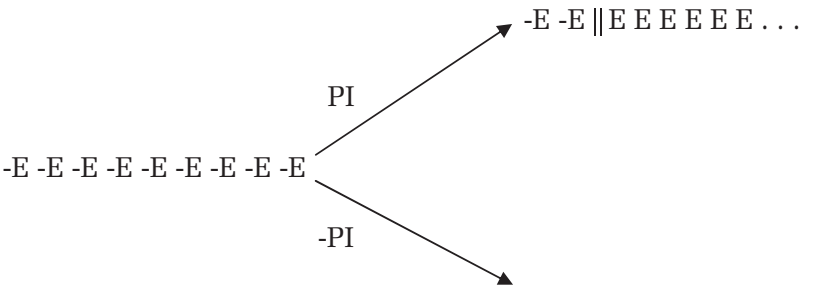


Figure 12.4. Unilateral Endorsement, Because the -PI Branch Is Truncated: Example of Lee the Soccer Player.

the premature termination of the lower -PI branch. Typically, the loss of the ability to make endorsements is due to death or serious mental impairment.

Consider an example of the kind of case illustrated in figure 12.4: Lee is a 25-year-old soccer player. He fell in love with soccer when he was in elementary school. He has built his life plan around playing and eventually coaching soccer. For years he has focused on playing for the U.S. World Cup soccer team. He is intensely competitive. He has often told his friends that if something were to happen to him to prevent him from playing soccer, he would rather not go on living.

Lee was recently selected to play on the U.S. World Cup soccer team. Shortly after, on his way home from practice, he was involved in an auto collision that left him permanently paralyzed from the waist down. Having been deprived of all hope of realizing his dreams, Lee is no longer motivated to live. He wants to commit suicide. Can involuntary commitment to prevent him from committing suicide be justified?

This case is different from the example of feverish Arnold, who must be restrained from stepping out of a third-story window, because, in this case, the hypothetical endorsement of intervention standard is satisfied. It is clear that Lee's decision to commit suicide is his stable judgment about how to further his settled values and preferences and he would not endorse intervention to prevent him from committing suicide in this sort of case.

Nonetheless, the case is not as clear-cut as the backward-looking hypothetical endorsement of intervention standard would imply. To see why not, suppose that if Lee is prevented from committing suicide, it is almost certain that he will be able to make a life for himself involving new values and preferences that will, over time, become settled, so that at some point his stable judgment about how to further his settled values and preferences will favor his new life and endorse the intervention to prevent his earlier self from committing suicide.[18] Suppose statistical studies have shown that if they are prevented from committing suicide, the overwhelming majority of athletes like Lee who suffer career-ending injuries will go through a period of depression for 1 or 2 years, after which time they will put together a new life they regard as worthwhile and they will come to unequivocally endorse the intervention necessary to prevent them from committing suicide during the extended period of depression. To add to the force of the example, suppose that if he does not commit suicide, it is reasonable to expect that Lee will have 40 or more years of life, and during the last 38 years he will consistently endorse the intervention that prevented him from committing suicide.

This case is not one of future bilateral endorsement, because if Lee commits suicide, there will be no future self to endorse or fail to endorse anything. This fact is represented in figure 12.4 by the fact that the -PI branch of the diagram terminates prematurely. However, in this case, I believe suitably humane intervention to prevent Lee from committing suicide can be justified, and justified for the same kind of reason it is justified in cases of future bilateral endorsement. It seems to me it is reasonable to take Lee's future

judgment when he is prevented from committing suicide (the PI branch of the diagram in figure 12.4) to be more reliable than his current judgment in favor of suicide. His current judgment may be that it is not worth the pain of rebuilding his life to get to the point where he has new values to pursue. However, his later self will have lived through the transformation and will be in a better position to evaluate whether it was worth enduring. If so, then the import of the missing second branch in figure 12.4 is that there is no other equally reliable judgment that conflicts with the hypothetical later judgment endorsing the paternalistic intervention, so the intervention satisfies the most reliable judgment standard as soft legal paternalism.

A second complicating factor in the Lee example is that if the intervention occurs, Lee will undergo a transformation in values so radical it might seem misleading to think of the later Lee as the same person as the earlier one. If they are not the same person, then allowing the later Lee's judgments to overrule those of the earlier Lee about what is good for him would be objectionable in the same way that allowing the judgments of another person to overrule Lee's own judgments about what is good for him would be.

Advocates of continuity theories of personal identity (e.g., Parfit 1984; Regan 1983) would find this sort of response particularly compelling. This seems to me to be a mistake. Even large changes in evaluative beliefs about what constitutes one's own good can be appropriate, when the occasion warrants a large change. Part of being a person is to be capable of such transformations when appropriate. It seems to me that Lee's situation is just such an occasion. So it seems to me that Lee's situation is one in which it would be reasonable to expect one and the same person to undergo a large transformation in his evaluative beliefs about what constitutes a good life.

If I am right that the most reliable judgment standard would categorize a suitably humane legal policy of intervention to prevent Lee from committing suicide as soft legal paternalism, then it is not necessary for the target's future selves to bilaterally endorse intervention for it to be justifiable. Bilateral future endorsement is required when there are two equally reliable hypothetical future selves. However, when the -PI branch terminates prematurely, then the lack of a second branch makes the judgment on the remaining branch authoritative.

Even when the most reliable judgment standard justifies intervention to prevent someone like Lee from committing suicide, it will not justify intervention in perpetuity. Such intervention would have to be for a limited time, because the longer Lee's determination to commit suicide persisted, the less probable it would be that he would ever change his mind.

Against Euthanasia

In figure 12.4, the -PI branch of the diagram is truncated. Symmetry considerations suggest there is another kind of exception to the hypothetical

endorsement of intervention standard, when the PI branch is truncated (see figure 12.5).

Figure 12.5 represents a case in which paternalistic intervention renders the target incapable of autonomous judgment. Thus it represents paternalistic intervention that kills or severely mentally impairs the target of the intervention, which the target does not endorse at the time of the intervention. Although figure 12.5 may represent a logical possibility, I do not include it as a category of soft legal paternalism, because, I do not believe there are any realistic examples of policies that fit it.

To explain why not, I try to construct one. Consider, for example, euthanasia. I want to limit the example to cases of euthanasia in which the potential target's stable judgment at the time of the intervention is that being killed would *not* be good for her and on that basis does *not* endorse being euthanized. To fit the diagram in figure 12.5, there would have to be an example in which it is reasonable to believe that a potential target of euthanasia might object to being euthanized now, but if euthanasia were not performed, would later come to endorse having been euthanized earlier. How could there be such an example? It is always possible that sometime in the future the target might change her mind about being killed and ask to be killed. Perhaps she finally becomes convinced there is no other alternative to suffering a prolonged, painful death. Even then she would be endorsing only assisted suicide. She would not be endorsing euthanasia, because she would not be endorsing being killed against her will. Because I cannot think of any plausible scenario in which reasonable people would endorse a policy of being killed (or being severely mentally impaired) for their own good against their will, I do not believe there are any exceptions to the hypothetical endorsement of intervention standard that fit the diagram in figure 12.5.[19]

The cases discussed so far are exceptions to the claim that the hypothetical endorsement of intervention standard is a necessary condition for soft legal paternalism. I turn now to a briefer discussion of exceptions to its sufficiency.

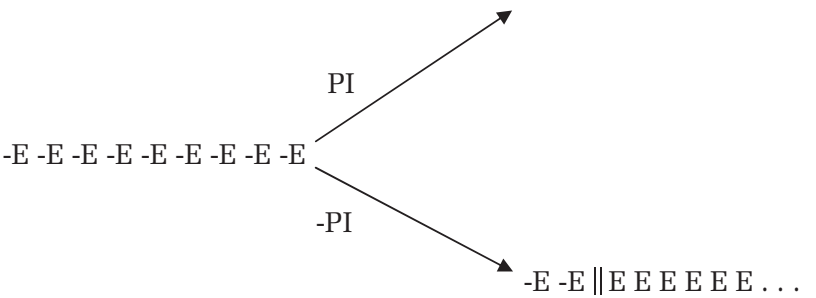


Figure 12.5. The PI branch Is Truncated; An Empty Category.

The Hypothetical Endorsement of Intervention Standard Is Not a Sufficient Condition for Legal Paternalism to Be Soft

Exceptions to the sufficiency of the hypothetical endorsement of intervention standard will be cases in which the target of paternalistic intervention endorses the relevant kind of intervention but the intervention is not justifiable. The previous discussion of cases of future bilateral endorsement (e.g., the example of Allen and the drug RD) immediately suggests a parallel class of cases of future bilateral antiendorsement, as illustrated in figure 12.6.

Figure 12.6 illustrates the dual of the example of Allen, illustrated in figure 12.1 above. Suppose Fred's stable judgment based on all the relevant information endorses paternalistic intervention to prevent him from taking drug HD. It would seem that such intervention would be soft legal paternalism, because it would satisfy the hypothetical endorsement of intervention standard. However, suppose, in addition, that good statistical evidence makes it reasonable to believe that if the intervention takes place, Fred's future self will come to unequivocally endorse nonintervention and that if the intervention does not take place, Fred's future self will also come to unequivocally endorse nonintervention. If we know that Fred's future selves bilaterally endorse nonintervention, the most reliable judgment standard would not endorse paternalistic intervention.

As before, talk of future selves is only a heuristic for talking about Fred's actual and hypothetical attitudes. The heuristic can be replaced by setting a threshold for the probability that a member of the target population of potential users such as Fred would come to unequivocally endorse nonintervention (UE-), if the intervention were to take place (Prob(UE-/PI)), and the probability that a member of the target population of potential users

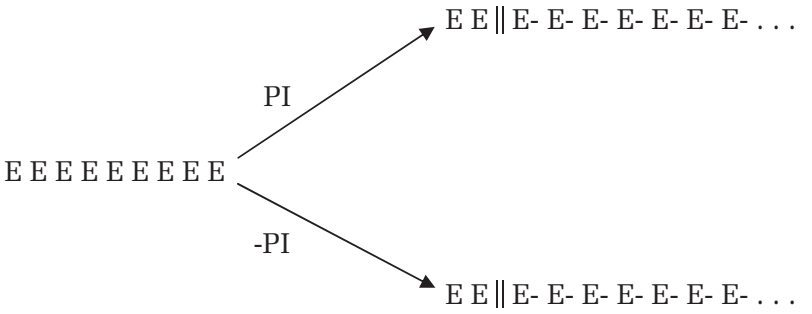


Figure 12.6. Bilateral Future Antiendorsement of Paternalistic Intervention that the Earlier Self Endorses. The target of the potential intervention would endorse it (E) at the time of the potential intervention, but it is reasonable to believe the target would come to endorse nonintervention (E-), both if the intervention were to take place and if it were not to take place.

such as Fred would come to unequivocally endorse nonintervention, if the intervention were not to take place (Prob(UE-/-PI)). For reasons discussed above, I believe the default threshold value for these probabilities should be 0.5.

Continuing the symmetry with the earlier discussion of necessity, there are even some exceptions to the sufficiency of the hypothetical endorsement of intervention standard involving only unilateral future nonendorsement. They are the cases illustrated in figure 12.7.

Figure 12.7 illustrates a kind of case very similar to the example of Lee the soccer player, illustrated in figure 12.3 above. Albert wants to commit suicide but lacks the nerve to do it. Albert's stable judgment about how to further his settled values and preferences endorses the state's assisting his suicide. In this case, the PI branch is truncated, because if the state helps him to commit suicide, there will be no future self to endorse anything. Suppose there is good reason to believe that if the state does not assist his suicide, Albert will come to unequivocally endorse not intervening in such cases. This case is more clear-cut than the example of Lee the soccer player. It would be hard paternalism for the state to assist Albert to commit suicide.

There is one further logical possibility, a case with a truncated–PI branch. This kind of possibility is illustrated in figure 12.8.

I do not believe there are any realistic examples of policies fitting figure 12.8. To see why, I try to construct one: Florence will die unless there is paternalistic intervention to keep her alive. Her current self endorses the intervention. However, later in her life she will become miserable and she will unequivocally judge that she would be better off dead. Could she come to endorse a policy of allowing her to die at the earlier time, when she wanted to be kept alive? I don't see how she could reasonably do so. The only reasonable policy for her to endorse is the policy of keeping her alive when she wants to be kept alive and not keeping her alive when she does not. So I do not see how figure 12.8 could generate realistic examples of soft paternalism.

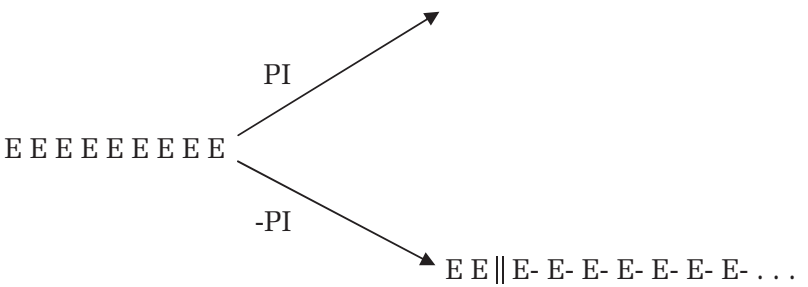


Figure 12.7. Unilateral Future Antiendorsement, Because the PI Branch Is Truncated: The Example of Albert.

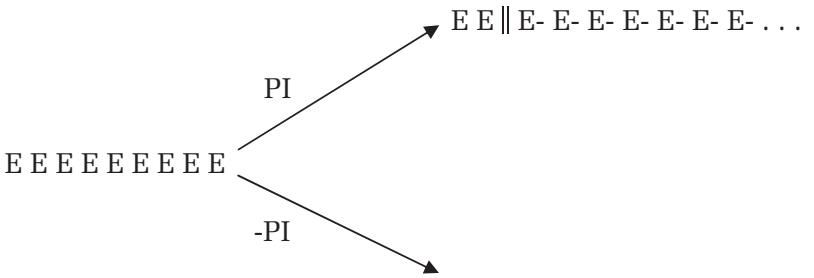


Figure 12.8. Unilateral Future Antiendorsement, Because the -PI Branch Is Truncated: An Empty Category.

The Most Reliable Judgment Standard for Soft Legal Paternalism (Final Version)

The guiding idea of the most reliable judgment standard is that legal paternalism is soft when it is reasonable to believe it is part of a policy that would be endorsed by the target's most reliable judgments or when it is reasonable to believe it is part of a policy the target would voluntarily endorse at the time of intervention and the target's most reliable judgments would not oppose the policy. On the assumption that a person's later judgments are generally more reliable than her earlier ones, the most reliable judgment standard is a forward-looking standard that explains why no backward-looking standard, such as the hypothetical endorsement of intervention standard, is adequate.

With the addition of some qualifications that I explain in notes, I am now in a position to give the final statement of the most reliable judgment standard:

Most Reliable Judgment Standard for Soft Paternalism—Final Version. In a case in which no minority of the target population will be severely disadvantaged by a policy of legal paternalism PI,[20] the policy of PI toward a target population is soft whenever the following condition is true of the target population and there is no sub-population of the target population that can be reliably and at reasonable cost distinguished from the target population for which it is not true:[21]

(a) the target would voluntarily endorse the relevant policy of intervention at the time of intervention and neither of the following two exceptions holds: (i) [Future Bilateral Majority Anti-Endorsement] it is reasonable to believe that, for the target population, both Prob(UE-/PI) and Prob(UE-/-PI) are greater than the threshold value for soft paternalism (typically 0.5) or (ii) [Future Unilateral Majority Anti-Endorsement, Because the PI Branch is Truncated] it is reasonable to believe that the PI branch will be truncated and Prob (UE-/-PI) for the target population is greater than the threshold value for soft paternalism (typically 0.5); or

(b) the target would not voluntarily endorse the relevant policy of intervention at the time of intervention, but one of the following two exceptions holds: (i) [Future Bilateral Majority Endorsement] it is reasonable to believe that, for the target population, both Prob(UE/PI) and Prob(UE/-PI) are greater than the threshold value for soft paternalism (typically 0.5); or (ii) [Future Unilateral Majority Endorsement Because the -PI Branch is Truncated] it is reasonable to believe that the -PI branch is truncated and Prob(UE/PI) for the target population is greater than the threshold value for soft paternalism (typically 0.5).

I have illustrated the most reliable judgment standard by supposing that the relevant conditional probabilities were based on statistical evidence. The standard does not require statistical evidence, but that is usually the best kind of evidence to have. It is unfortunate that there is not more statistical evidence of this kind. It would be very useful for young people choosing careers to know what percentage of those who chose the various careers they are considering are satisfied with their choice when they retire. Similar statistics would be useful for other important life decisions. What is the optimal number of children to have? Find out what percentage of those who had zero, one, two, three, and so forth would do the same again. What is the ideal age to marry? Find out what percentage of those who married at each age would marry at the same age again. There is a great potential for generating lots of useful statistical information of this kind.

Suppose that a government that follows the most reliable judgment standard is contemplating a paternalistic ban on certain designated activity. If statistical evidence of the target population's attitudes toward a ban in both conditions, with the ban and without it, were required before a ban could be justified, then no new ban could ever be justified, because, before the ban was implemented, there could not *be* any statistical evidence of what their attitudes toward the ban would be after it was implemented. A more typical case will be one in which there is no ban and the need for one is so great that a large majority of the target population favors it. If the reasons for a ban are salient enough and the majority support is large enough, it is reasonable for the advocates of a ban to project that even after the ban has been implemented, a majority of the target population will still endorse the ban. However, once the ban has been implemented, it would still be necessary to gather data to find out what the target population's attitudes toward the ban actually were. This is particularly important, because the ban may have had undesirable side effects that were not anticipated before it went into effect.

Ideally Reliable Judgments

The most reliable judgment standard justifies overruling a person's less reliable earlier judgments on the basis of what it is reasonable to expect her more reliable later ones to be. Once hypothetical judgments can be entertained, why not entertain even more reliable hypothetical judgments than

the judgments people can ever make? For example, undoubtedly people's judgments about what is good for them would be more reliable if they were omniscient. Omniscience would be the highest ideal of reliability. In evaluating a policy of legal paternalism, we could ask, would the target endorse the policy if she were omniscient? If so, because this hypothetical endorsement would be more reliable than her current or future judgment, the most reliable judgment standard would justify using it to overrule the target's current and future judgments.

There are two problems with this sort of suggestion when evaluated by the main principle. The first is the obvious one that we are not omniscient so we have no way to know what the target's judgment would be if she were omniscient. This objection is not decisive, because it might have been the case that fallibly trying to apply this standard would have been a good way of promoting people's life prospects. Thus, the second objection is the most important one. The second objection applies to all ideally reliable judgment theories. The track record of such theories is abysmal. Berlin (1969) has reminded us of the awful things that have been done to people on the grounds of an idealized theory about what is good for them. Idealized theories are not good ways of promoting the life prospects of the targets of intervention, because they are vehicles for overriding judgments the target herself makes or is expected to make with hypothetical judgments the target never could make. No human being could ever be omniscient or otherwise ideally rational. Because the policy of basing paternalistic intervention on such theories is such a disastrous one, it would not be endorsed by the main principle. For this reason, ideally rational judgment theories cannot be used to justify legal paternalism. The only hypothetical judgments that can justify paternalistic intervention in the acts of an autonomous target are the judgments the target herself, with all her cognitive and other constraints, would make or would probably make, if she had the opportunity to do so.

The Most Reliable Judgment Standard Is Not an Invitation to Legal Paternalism

Because my main focus in this chapter has been on the categories of soft legal paternalism that cannot be accounted for by the hypothetical endorsement of intervention standard, I may have given the mistaken impression that the most reliable judgment standard classifies lots of legal paternalism as soft. Actually, very little legal paternalism qualifies as soft paternalism under the most reliable judgment standard. There are two reasons for this. First, our later selves are typically very conservative about endorsing policies of paternalistic government intervention in our choices. Even when they rue what they take to have been a mistaken earlier decision, most people judge it is better that they be allowed *to make mistakes and to learn from them* than that the government interfere to prevent them from making what they will

later judge to have been a mistake. It is usually only in cases in which mistakes can be expected to cause severe, unavoidable losses that most people would endorse legal intervention to be prevented from making them.

The second reason very little legal paternalism qualifies as soft paternalism under the most reliable judgment standard is that, in most cases, a person's current judgment about a policy of paternalistic intervention is the best evidence we have of what her future judgment will be. It is only in unusual cases that we have good reason to believe that her future judgment will overrule her current one. However, I think it would be valuable to collect more evidence of this kind.

The Most Reliable Judgment Standard Fits the Legal Standard of Autonomy

What is autonomy? In the first volume, I introduced a consequentialist conception of autonomy as a combination of good judgment (the ability to make reliable judgments of one's own good) and self-determination (to have one's actions be based on one's judgments) (Talbott 2005, 131–132).

Why does the main principle endorse the adoption of this standard for autonomy? In the first volume, I emphasized the importance of governments receiving reliable feedback on the extent to which their policies promote life prospects (2005, chap. 6). In chapter 8 of this volume, I have added the Millian consideration that each of us, in living our life, is conducting an experiment to determine the best life for human beings. These experiments in living are the main source of progress in every area of life—for example, in science, medicine, technology, literature, music, art, philosophy, religion, food, and fashion. Even those who are not very experimental benefit from the experiments of others.

What is the main principle's criterion for declaring individuals to be competent to conduct experiments in living? It is the same criterion as for declaring individuals to be competent to enter into contracts. They must be autonomous in my consequentialist sense. For those who have self-determination, autonomy depends on the degree of reliability of their judgments about what is good for them. When legal systems define a default age for legal competence what they are really defining is a threshold level of reliability of people's judgments about what is good for them. The main principle favors setting the threshold at the youngest age at which general life prospects are higher if minors are released from their minority at that age than if not. At this threshold, it is better for life prospects generally if individuals are given the legal power to bind themselves on the basis of their own judgments about what is good for them than they would be if other people—their guardians, typically their parents—retained the ability to legally bind them on the basis of the guardians' judgments about what is good for them.

Imagine a graph of the average reliability of people's judgments about their own good as a function of age. It starts off near zero at birth and increases with age. At some point—say at age 18 or 21—for the first time, the average reliability is high enough that general life prospects will be higher if people at that age are assumed to be legally competent than they would be if the default age of competency were set later. At that point, substituting other people's judgments about what is good for them (even their parents') will generate lower life prospects than giving them the power to act on their own judgments. That is the reliability threshold that determines legal autonomy or competence.

What is the legal standard for autonomy? There are various legal standards. Consider the standard for appointment of a legal guardian. A legal guardian is typically appointed when an adult is unable to care for his own physical health or to manage his financial affairs. These are clearly attempts to articulate a minimal standard of reliability of judgments concerning one's own good.

Another standard is the legal test for incompetence used in involuntary commitment proceedings. People can be involuntarily committed if they are a threat to themselves or to others. It may not be obvious, but this is a test for autonomy, in my consequentialist sense. The first half of the test, threat to oneself, fits the consequentialist conception exactly. There are two ways one could be a threat to oneself: first, lacking good judgment, one would not know what was good for one, and, second, having good judgment, but lacking self-determination in my sense, one's actions would not be based on one's judgment of what was good for one. What about the second half of the test, threat to others? Here again, all that is required is to be autonomous in the sense captured by my consequentialist conception. First, one needs to be reliable enough to form accurate beliefs about the legal penalties for harming others. Given this understanding, there are two ways that being a threat to others can be the basis for involuntary commitment: first, lacking good judgment, one might not be able to understand why punishment would be bad, and, second, lacking self-determination, one might not be able to stop oneself from harming others even though one wanted to avoid punishment. Of course, there is a third way that punishment fails to deter: when people with good judgment and self-determination choose to break the law and try to avoid being apprehended. Those people are not involuntarily committed as incompetent. They are tried as criminals and punished.

What is nonconsequentialist autonomy? This is a deep puzzle. Kant [1785] thought that it was the ability to make choices that were not causally determined. He thought we had no way to know whether anyone's choices were autonomous in this sense, not even our own. In our own case, he thought we just had to assume that we had it. Or to be more precise, I have to assume that I have it. I have no way of ever knowing whether you do or not.

The obvious question for such a view of autonomy is, how could we ever have a legal test for it, if we can't even know whether we have it ourselves?

One reply might be that if you think you have it, then you have it. This cannot be right. There are many mentally ill people whose choices are products of their illness, though they don't realize it.

There is in the law of criminal liability one remnant of the nonconsequentialist conception of autonomy. In criminal law, the insanity defense is a bar to criminal liability if the defendant did not know right from wrong. This test fits well with the Kantian conception of autonomy, but it is a complete puzzle to me why it should be thought to have anything to do with criminal liability. Suppose that a sociopath is put on trial for murder and it is determined that he committed the crime, that he is autonomous in the consequentialist sense (good judgment and self-determination), that he knew that if he were caught he would face the possibility of life imprisonment, that he thought he was clever enough that he could avoid being caught, but that he does not know right from wrong. How could his not knowing right from wrong make a difference to whether or not he should be punished? He killed another person knowing that the penalty was life imprisonment but thinking that he could avoid getting caught. How could we expect punishment to deter crime unless people like him are punished when they are caught?

If knowing the difference between right and wrong is irrelevant to legal liability, then it seems that the nonconsequentialist notion of autonomy is dispensable in the law. The nonconsequentialist will reply that it is not dispensable to our judgment that a person *deserves* to be punished. This is too large a topic for me to take up here. However, my example of the sociopath illustrates why the main principle endorses a legal standard that punishes sociopaths for their crimes, whether they have nonconsequentialist autonomy or not.

From Autonomy to the Most Reliable Judgment Standard

The main principle determines a legal standard of competence or autonomy based on a threshold of minimal reliability of their judgments of their own good. This threshold represents the point at which, as a matter of policy, if people have the power to legally bind themselves and to conduct their own experiments in living, life prospects will be higher than if other people—their guardians, typically their parents—have that power. Their lives and other peoples' lives will generally go better if they have that power than if other people have that power. This creates a presumption that other people's judgments of what is good for them should not be the basis of limiting their autonomy.

However, this legal threshold is not the termination of the process by which one's judgments of one's own good become more reliable. The process does not end at 18 or 21; it continues. Although there is no guarantee that one's judgments about one's own good will generally become more reliable with age, they generally do so, far beyond 18 or 21, until a relatively late age

at which our capacities may decline. Even if a policy of allowing other people's judgments to overrule an autonomous agent's own judgment about what is good for her would not generally promote life prospects, there is another policy that would do so: the policy of allowing an autonomous person's own later, more reliable judgments to overrule her earlier less reliable ones. Thus, the main principle endorses the most reliable judgment standard as a ground-level standard for defining the scope of the right against legal paternalism.

Conclusion

The most reliable judgment standard provides the consequentialist with a ground-level standard for distinguishing between soft and hard legal paternalism. In the next chapter, I use that standard to define liberty rights against legal paternalism as human rights.

THIRTEEN

Liberty Rights and Privacy Rights

> The principle requires liberty of tastes and pursuits; of framing the plan of our life to suit our own character; of doing as we like, subject to such consequences as may follow; without impediment from our fellow-creatures, so long as what we do does not harm them even though they should think our conduct foolish, perverse, or wrong.
>
> —J. S. Mill

In the previous chapter, I explained the most reliable judgment standard for soft legal paternalism and gave the consequentialist rationale for employing it as a ground-level moral principle to distinguish soft (permissible) from hard (impermissible) legal paternalism. In this chapter, I specify the contours of the rights against hard legal paternalism more precisely. Also in this chapter, I consider the question of whether there are any privacy rights that should be universally protected as human rights.

The Evolution of a Liberty Right against Legal Paternalism

Perhaps the most remarkable development in U.S. Supreme Court jurisprudence since the mid-twentieth century has been the development of a constitutional right that is not to be found in the U.S. Constitution. It has taken almost 50 years for the Court to clearly articulate what kind of right it is, though they have a long way to go to fully implement it. Not only does this right not appear in the U.S. Constitution, it does not appear by name in *any* human rights document anywhere in the world, though many of its instances do. Though the right was first articulated as a right to privacy, it is really a liberty right against legal paternalism. In this chapter I begin by discussing the liberty right and postpone the discussion of privacy rights to the end.

Why would I think that there should be human rights against legal paternalism? There is no general right against legal paternalism recognized anywhere in the world. In the first volume (Talbott 2005) I suggested that the entire history of human rights is a history of rebellion against paternalistic rationales for oppression: the belief the that commoners needed a monarch to look after their interests; that colonials needed the colonists to look after their interests; that slaves needed a master to look after their interests; that women needed a father and then a husband to look after their interests; that people

with disabilities needed custodians rather than the removal of the barriers that prevent them from living independently. But even if human rights are a response to paternalistic oppression, it does not follow that there should be a general right against legal paternalism. Perhaps there should be a right only against *oppressive* legal paternalism. To decide the question, it is useful to start with a brief history of the development of liberty rights against legal paternalism.

The most important event in the historical development of liberty rights against legal paternalism was the development of a right to religious freedom. The reason is simple: There is no greater harm a person could do to herself than to bring it about that she suffers unbearable torment for all eternity. Suppose I believe that will be your fate if you do not practice my religion. I propose to save you (and others like you) from eternal suffering by making it illegal for you to practice any religion but mine. This legal establishment of my religion would be an example of legal paternalism, because enforcing it would involve my overruling your own judgment about what is good for you.[1] A right to freedom of religion represents a rejection of this kind of paternalism. Once it is allowed that people should be free to make and follow their own judgments of what will be to their eternal benefit and harm, it is hard to see why they should not be equally free to make and act on less momentous decisions about what is good for them.

After a right to religious freedom, the second most important step in the development of a right against legal paternalism is the development of the rights that guarantee the necessary background conditions for autonomy, especially the rights to civil liberties. In addition to being essential background for autonomy, rights to freedom of expression, freedom of association, and freedom of the press are important steps in the development of rights against legal paternalism, because they involve a recognition that those in authority should not be deciding which ideas it is bad for people to be exposed to or to think about or to discuss.

With all these rights in place, there are many potential pathways to further rights against legal paternalism. The typical pathway involves the recognition of other choices that, like the choice of a religion, are deeply personal choices, the effects of which are borne primarily by the person making the choice. In spite of opposition from almost all major religions and in spite of laws to the contrary, in the United States, a right against legal paternalism has gradually developed around personal choices concerning sex, reproduction, and death.

Freedom from Paternalistic Interference in Choices Concerning Sex and Reproduction

The main legal development has been the U.S. Supreme Court's articulation of a right not explicitly found in the U.S. Constitution, first introduced as a

"penumbral" right to privacy. The leading case in the development of the right to privacy was *Griswold v. Connecticut*,[2] in which the Supreme Court declared unconstitutional a law forbidding the use of contraceptive drugs or devices. Declaring that the Bill of Rights creates a penumbra that includes a right to privacy, the Court overturned the statute as an unconstitutional limitation on a right to marital privacy.

Because the statute in *Griswold* held out the possibility of searches of a couple's bedroom for contraceptive devices, it was reasonable to think the implicated right was a right of privacy. In fact, the issue raised by the statute was whether people should be free to make certain sorts of decisions and to act on them without the government overruling their judgment.[3]

Two years later, the Supreme Court expanded the protection against legal paternalism when it struck down as unconstitutional a Virginia law against miscegenation in *Loving v. Virginia*.[4] Though the law was struck down on the grounds that it involved invidious racial discrimination that could not survive strict scrutiny, this conclusion was difficult to sustain, because the law treated both races equally. In its decision, the Court also articulated what would turn out to be the most compelling basis for the decision, the protection of an important sphere of personal autonomy: "The freedom to marry has long been recognized as one of the vital personal rights essential to the orderly pursuit of happiness by free men."[5]

It was reasonable to think the right articulated in *Griswold* was a right to privacy, because it concerned decisions and actions by married couples in their bedroom. In *Planned Parenthood v. Casey* it was properly categorized as a liberty right, protecting "the most intimate and personal choices a person may make in a lifetime, choices central to personal dignity and autonomy."[6]

In 1986, the Supreme Court refused to accept a clear implication of its own prior decisions, when, in *Bowers v. Hardwick*,[7] it upheld the constitutionality of a Georgia prohibition of sodomy. The Court upheld the prohibition though there could hardly be a more intimate and personal choice than the choice of a sexual partner made in the privacy of one's own bedroom. Ironically, 12 years later the Georgia Supreme Court did rule the Georgia sodomy law unconstitutional on the basis of the same kind of reasoning that the U.S. Supreme Court had used in *Griswold* and *Planned Parenthood v. Casey*.[8] It took 5 more years for the U.S. Supreme Court to accept the implications of its own earlier decisions, by overruling *Bowers* in *Lawrence v. Texas*.[9] In *Lawrence* the Court defined the sphere of protected liberty to include "personal decisions relating to marriage, procreation, contraception, family relationships, child rearing, and education."[10]

In his dissenting opinion in *Lawrence*, Justice Scalia pointed out that a clear implication of the *Lawrence* decision is that prohibitions on same-sex marriage are unconstitutional. He was right. Unfortunately, it seems unlikely the Supreme Court will acknowledge this implication anytime soon. Thus, in matters of sex and reproduction, Europe and Canada have now surpassed the

United States in the working out of a sphere of personal autonomy free of paternalistic government interference.[11]

Freedom from Paternalistic Interference in Choices Concerning One's Own Death

In 1976 the New Jersey Supreme Court interpreted what was then thought of as a constitutional right to privacy to include a right to refuse extraordinary measures to save one's life and a right of a guardian to refuse them for an incompetent patient.[12] In 1990, the U.S. Supreme Court extended the right to include a right to refuse ordinary life support.[13] When a patient was incompetent, the Court required clear and convincing evidence of intent. As a result of these decisions, normal adults in the United States can exercise their right to refuse extraordinary measures or to refuse life support by executing an advance directive or living will. Because the decisions to refuse extraordinary measures or to terminate life support are typically made in a hospital, not in the privacy of one's own home, and because they typically involve interactions with strangers, these cases show even more clearly the right to "privacy" is a misnomer. It is really a right to a sphere of personal autonomy free of legal paternalism.

In 1997, the U.S. Supreme Court declined to extend this right to include a right to assisted suicide, even though an *amici curiae* brief was filed in support of such a right by six of the most prominent political philosophers of the last half of the twentieth century.[14] The six philosophers invited the Court to take the step of explicitly announcing a right to a sphere of autonomy protected from legal paternalism that would include a right to assisted suicide at the end of life. Even though the Court did not accept the invitation, five of the justices made it clear that their decisions in the instant cases did not foreclose the larger constitutional question. The momentum for such a right continues to build. Oregon voters have twice endorsed it and the Supreme Court has upheld their endorsement.[15] In 2008 the state of Washington also adopted such a law.

I believe the six philosophers were correct to urge the Court to reconceptualize the liberty rights against legal paternalism to include an end-of-life right to assisted suicide. Nonconsequentialists who value autonomy would tend to agree. The challenge to the nonconsequentialists is to explain why the right to assisted suicide would be limited to end-of-life choices. I discuss this issue shortly.

Another Step in the Evolution of a Right against Legal Paternalism

Another important development in the evolution of a human right against paternalism has been the adoption and ratification of the U.N. Convention on

the Rights of Persons with Disabilities. Here is the U.N.'s description of the significance of the convention:

> The Convention marks a "paradigm shift" in attitudes and approaches to persons with disabilities. It takes to a new height the movement from viewing persons with disabilities as "objects" of charity, medical treatment and social protection towards viewing persons with disabilities as "subjects" with rights, who are capable of claiming those rights and making decisions for their lives based on their free and informed consent as well as being active members of society. (U.N. 2009)

The convention represents a transformation in the understanding of disability that has taken place over the past 50 years or so. The convention creates a presumption that persons with disabilities are to be included as full and participating members of the community and to be regarded as bearers of the full human rights. Implementing this convention will go a long way toward eliminating unwarranted legal and cultural paternalism toward persons with disabilities.

A Social Framework for Good Judgment

In the first volume I emphasized the importance of the social framework for the individual's development of good judgment (i.e., the ability to make reliable judgments of one's own good). I also emphasized that in my sense of the term, autonomy, understood as the combination of good judgment and self-determination, is a social achievement. The most important social conditions for the attainment of individual autonomy are the basic human rights. They are social conditions, because, for each of us, our individual autonomy depends on other people's having those rights (Talbott 2005, 136–137).

A minimal condition of adequacy on my consequentialist account of a right against legal paternalism is that when the social framework for autonomy is established, citizens who develop their capacities in that framework come to judge that it is good for them to do so. If they concluded that it was not good for them to be autonomous, then my consequentialist case for a right against paternalism would collapse. This is a theoretical possibility, as I described in the first volume (Talbott 2005, 185–186). Fortunately, the situation in the actual world is not so dire. In the actual world, autonomous citizens typically voluntarily endorse having a social framework that makes them autonomous.

This sort of voluntary endorsement explains why many government policies that are described as *paternalistic* raise no concern of unjustified or hard paternalism. The reason is that they are policies that a majority voluntarily endorses (or would endorse, a qualification that I leave implicit in the following discussion) to limit their own choices.

This potential for people to voluntarily favor paternalistic restrictions on their liberty is illustrated by Odysseus, who had his crew stop their ears (but not his) and tie him to the mast of his ship, so that he could hear the song of the Sirens without being able to steer the ship toward the Sirens and certain destruction. Odysseus ordered the crew to refuse to untie him no matter how hard he struggled to free himself. The crew did so. On any reasonable account, consequentialist or nonconsequentialist, the crew's paternalism toward Odysseus is soft paternalism. Such paternalism I refer to as a *solution to an Odysseus problem*. There are many more examples of the paternalism of this kind.[16]

Legally Enforced Expert Opinion

In chapter 10 I discussed how consumers improve the reliability of their product judgments by relying on experts, such as those at *Consumer Reports*. There the reliance was voluntary. However, there are cases in which the government gives expert judgment coercive force and still the paternalism is only soft, because they are solutions to an Odysseus problem.

Thus, for example, news stories about a toxic food or hazardous toys or defective tires typically generate widespread questions from the press and public about why government regulators did not effectively protect the public from the hazardous product. This is clear evidence of majority endorsement of safety regulation. So it is not hard paternalism. Even though there are undoubtedly some people who would rather not have any government safety regulations, the laws do not constitute hard paternalism toward them. This is an example of the *majority spillover effect*. Such laws are not hard paternalism so long as the majority favors the laws to promote their own good, not because they think they are good for those in the minority who don't endorse them. There is nothing wrong with the majority *thinking* that a paternalist law is good for those in the minority; no one has a right that other people not *think* they are making a foolish decision. The law becomes hard paternalism only when the reason for adopting it is for the good of the minority who does not endorse it (when it is not reasonable to believe that they would in the future either).

Scanlon (1972) discusses the possibility that we might even delegate to the government the authority to protect us from our own bad reasoning. As Scanlon points out, this delegation would require safeguards not to compromise one's autonomy. Nonetheless, we can easily recognize it as an Odysseus problem in which the government gives effect to our more reliable judgment. This is the reason that government is permitted to prohibit pyramid schemes and to regulate speculative bubbles. Human beings find it almost irresistible to get caught up in such schemes when they see that all the participants are getting rich. This is one kind of bad reasoning that almost everyone is susceptible to.

The Main Principle and Legal Paternalism and the Example of Mandatory Retirement Savings

Thus far, I have been working out the elements of a consequentialist standard for soft legal paternalism, the most reliable judgment standard, as part of the pure theory of legal paternalism, in which undesirable side effects can be ignored. What is the relation between this pure theory and the main principle? For the reasons I explained in the previous chapter, the main principle endorses the most reliable judgment standard as a ground-level principle for the courts to apply to judge whether to uphold paternalistic laws. It is the ground-level principle endorsed by the main principle to define the contours of the right against legal paternalism.

The most reliable judgment standard does not tell us when a paternalistic law should be passed by the legislature. That decision will depend on a variety of considerations, including the cost of enforcement and the probability and severity of negative unintended consequences. But if the legislature does adopt a paternalistic law, the most reliable judgment standard provides the criterion by which the courts can judge whether or not the law should be upheld. The courts will not typically review the legislature's determination of the costs of enforcement and the potential negative side effects. The courts will focus their determination on whether or not the law constitutes soft legal paternalism under the most reliable judgment standard. If not, the law should be invalidated.

The most important difference between the most reliable judgment standard and the various autonomy-based principles of soft legal paternalism in the literature (e.g., the hypothetical endorsement of intervention standard discussed in the previous chapter) is that the autonomy-based principles are backward-looking, while the most reliable judgment standard is forward-looking, because it gives priority to a person's later, more reliable judgments about what is good for her. Because it is forward-looking, the most reliable judgment standard can help to resolve a number of puzzles about justified legal paternalism.

Consider, for example, the Social Security system in the U.S. as representative of a system for mandatory retirement savings. In chapter 11 I explained why such a system is not paternalistic, if it is an attempt to provide a social floor without holes. But even if it were enacted paternalistically, we can ask whether it would be soft or hard legal paternalism.

To answer this question, we must set aside those enrollees who would voluntarily enroll and would voluntarily make regular contributions to their retirement account with or without the law, because it does not matter to them whether or not the system is mandatory. Let's focus our attention on those who, if the system were not mandatory, either would not enroll or, if they did enroll, would not reliably make contributions to their retirement account. Let's look at their attitudes at two times, early in their careers and at retirement. Early in their careers, some of these people would probably

recognize that they were weak-willed and would welcome a precommitment device to make retirement savings mandatory. For these people, mandatory Social Security would be a solution to an Odysseus problem, and it would qualify as soft legal paternalism on any reasonable standard.

The more interesting cases are those people for whom, early in their careers, mandatory Social Security is *not* an Odysseus problem, because their stable judgment is that mandatory Social Security would not be good for them. On the nonconsequentialist hypothetical endorsement of intervention standard, making Social Security mandatory for these people would be hard paternalism. However, this seems to me to be a mistake. If, as I believe, at the time of retirement, a large percentage of these people will have changed their opinion and will have come to endorse Social Security's having been mandatory for *them*, then it is quite plausible that mandatory Social Security would obtain future majority bilateral endorsement, and thus qualify as soft paternalism under the most reliable judgment standard.[17]

We are now ready to investigate the proper scope of a human right against paternalism. Even if the most reliable judgment standard is the correct standard for legal paternalism, some issues, such as making mandatory motorcycle helmets or seat belt use, raise issues of paternalism that are too marginal to address with a human right. So I limit myself to more substantive issues of legal paternalism.

It is useful to begin with a recap of the liberty rights against paternalism that are now generally accepted:

1. A right to religious freedom. This is the first right against paternalism.
2. A right to sexual freedom. This is the right that the U.S. Supreme Court established in *Lawrence v. Texas*.
3. A right to reproductive freedom. This right would include contraception and some kind of abortion, though I don't see any decisive consideration about where to draw the line between permissible and impermissible abortions. The problem is that at the point at which the fetus has its own rights, the issue is no longer one of paternalism. This is an issue that is still in the process of being worked out. In *Roe v. Wade*, the U.S. Supreme Court hoped to settle the issue by drawing the line at viability. As medicine makes it possible to keep fetuses alive at earlier and earlier stages of development, the line defined by this decision is being altered.
4. A right to refuse medical treatment, including a right to refuse extraordinary care and to be removed from life support. This right is now well-established. However, I outline the basis for a possible exception to it when I discuss forced medical care.

I now turn to other potential extensions of the right against legal paternalism.

Right to Same-Sex Marriage

A right to marry is included in the U.N. Universal Declaration, but the full right would include same-sex marriage rights. Out of respect for the right to religious freedom, no religion should be forced to perform same-sex marriages, but no religion should be forbidden to perform them either. So long as public officials were available to perform same-sex marriages, the right would be protected.

In his dissent in the *Lawrence* case, Justice Scalia quite correctly pointed out that the logic of *Lawrence* would also apply to same-sex marriage.[18] That is why it is almost inevitable that someday the Court will recognize such a right. He also included protections for bigamy and adult incest as additional implications of the *Lawrence* decision. Is this correct?

The logic of *Lawrence* clearly does not apply to adult incest, for two reasons. The first is the potential for producing children with genetic defects. The second is perhaps even more important. Recall that the main principle evaluates practices, not individual acts. The practice of allowing adult incest would almost surely have, as a side effect, making young girls more prone to sexual abuse within the family. Sexual abuse of young girls within the family is a great cause of severe emotional and psychological problems that greatly affect life prospects. The main principle will not endorse a practice that would inevitably produce a significant increase in the sexual abuse of young girls.

Bigamy, or more generally polygamy, is a more apposite case. Mill opposed what he regarded as the paternalistic opposition to Mormon polygamy ([1859], 103–104). However, he would definitely have opposed what was also a part of the Mormon practice, marrying young girls at age 12 or 13. If child marriage is eliminated, should polygamy, understood to include both polygyny and polyandry, be permitted? Let's ask the question slightly differently: If the human rights I have advocated are legally protected, when all spouses are competent adults and agree to the arrangement, should multiple husbands or multiple wives be permitted? I don't see why not. When the human rights are legally protected, I do not believe that there will be many polygamous marriages, if any, because it will be rare that all parties will agree. It is no coincidence that polygamous cultures are cultures in which women are powerless. If they had equal rights, I believe that very few women would put up with the practice. Nor would very many men consent to polyandrous marriages.

The main reason for prohibiting polygamous marriage is that allowing it would have the tendency to communicate to girls that they are much less valuable than boys. This is a serious concern that makes me think that although the main principle would endorse a right to polygamous marriage in an ideal society, it may well not endorse it in the actual world until women have achieved economic equality with men.

Right to Suicide and Assisted Suicide?

Earlier I regretted the fact that the U.S. Supreme Court had declined the invitation in the Philosopher's Brief to declare a right to some cases of assisted suicide. The most reliable judgment standard provides a framework for explaining which cases of suicide and assisted suicide should be protected.

When a person has a terminal illness and has no reasonable prospect of any future free of pain or greatly diminished consciousness, should she be permitted to decide to end her life or to obtain assistance in ending her life painlessly and with dignity? Because there is no reasonable prospect of a transformation into a future self who would come to endorse having been prevented from committing suicide, these are cases in which intervention to assist in ending her life would satisfy the most reliable judgment standard. It is important to emphasize that I am not advocating euthanasia in such cases; I am advocating only giving effect to the person's own most reliable judgment about what is best for her. The right would require safeguards to make sure that suicide was voluntary.

What about other cases of suicide? Nonconsequentialists such as Feinberg (1986, 143) and G. Dworkin (1972, 32) can justify temporary restraint to determine whether the decision is voluntary. But once the decision has been determined to be voluntary, they have no grounds for interfering with it. Feinberg suggests that suicide does not raise important questions for a theory of legal paternalism, because no law can prevent a person who wants to commit suicide from doing so (1986, 145). This ignores the fact that all states have laws for involuntary commitment of those who are a danger to themselves. Often medical personnel or acquaintances can identify a potential suicide before the attempt. In any case, not all attempts are successful, so it is possible to identify potential suicides after an attempt. At one time Feinberg believed that there should be a presumption that suicide is nonvoluntary (1971, 11), but he later gave up that presumption (1986, 127). Dworkin even believes that there should be a presumption that the choice is voluntary (1972, 32). So these nonconsequentialists would seem to be committed to some sort of general right to commit suicide.

As illustrated by my discussion of the example of Lee the soccer player in the previous chapter, the most reliable judgment standard provides another ground for legal paternalism, based on the probable judgment of the future self that the intervention was good for her. What percentage of those who are prevented from committing suicide go on to die of natural causes? Richard Seiden (1978) did a study of everyone he could locate who had been restrained from committing suicide on the Golden Gate Bridge during the period from its opening in 1937 to 1971. In 1978, 94% of the 515 subjects he was able to account for were either still alive or had died of natural causes.[19] Of course, it was still possible that some of those who were alive would eventually commit suicide. But the fact that all of them had been alive for a minimum of 7 years shows that it is not true that if someone is prevented from committing

suicide, they'll just find another way to do it. Technically, it would be better to have data on whether those who were prevented from killing themselves endorse the intervention, but it is almost certain that a large percentage of them, much greater than 50%, would endorse it. So the most reliable judgment standard would classify the intervention as soft legal paternalism. And therefore the most reliable judgment standard would favor a right against legal paternalism that covered a right to suicide and assisted suicide at the end of life, but not a general right to suicide.

Slavery Contracts and Religious Vows

Slavery contracts raise different issues from typical cases of suicide or assisted suicide. We have already discussed in chapter 9 the cases in which prohibiting slavery contracts is a solution to a CAP. That prohibition is not even paternalistic. But there are, in theory, cases of slavery contracts in which a prohibition would not be a solution to a CAP. Call these cases of *voluntary slavery*. Nonconsequentialists have a hard time with voluntary slavery. Feinberg (1986, 78, 83–87), Nozick (1974, 331), G. Dworkin (1983, 111), and Thomson (1990, 283) all hold that such contracts should be enforceable in principle, if made voluntarily.[20]

It might be thought that the issue is moot, because there are not any voluntary slaves. This is not quite right. A religious vow of obedience is the equivalent of a slavery vow. Lots of people take such vows. Of course, the law of contracts does not apply to vows or promises (though it is not at all obvious why, on a nonconsequentialist account, it should not; cf. Fried 1981). But it would be easy to turn religious vows into contracts. Those who take religious vows have all their needs provided for by their religious order. So let us suppose that religious orders offered new members contracts of perpetual obedience, where in return for agreeing to the contract an individual would be assured of an education and provision for all his needs. Suppose, also, that the contract specifically stipulates that the remedy for breach is an order of specific performance (i.e., an enforceable court order of perpetual obedience). Should such contracts be legally enforceable? If they were legally enforceable, members of religious orders who signed them would have the legal status of slaves.

A nonconsequentialist has a hard time arguing that people should not be allowed to contract themselves into perpetual religious obedience. However, the most reliable judgment standard would clearly endorse making those contracts unenforceable. A policy of nonenforcement would almost surely be endorsed by the later self, because later selves who wanted to continue to comply with the obligation of obedience would be free to do so, and those who wanted to be released from it, would be able to leave.

Would the most reliable judgment standard favor making *all* contracts unenforceable? After all, if a person enters into a contract and later decides it

was not a good idea, isn't her judgment more reliable later, so shouldn't she be able to void the contract? The most reliable judgment standard does not apply on a case-by-case basis. What would be the effects on the practice of contracting, if a party could unilaterally void the agreement at any time? As I discussed in chapter 9, I think it is possible to define a subclass of the win-win contracts, for which allowing a period during which a party could unilaterally void the contract would promote life prospects, by reducing the frequency of and incentive for one-sided contracts. But it is clear that a general policy of unilateral voiding of contracts would have disastrous effects on the practice of contracting. The main advantage of the practice of contracting is that it enables parties to rely on each other. A policy of unilateral voiding would remove the ground for reliance and thus eliminate most of the benefits of the practice of contracting. So our future selves would never endorse such a policy.

This is not true for voluntary slavery contracts or voluntary religious contracts of perpetual obedience. There is no loss to allowing people to be voluntary slaves as long as they want, without making it possible for them to be held against their will. So an issue that is perplexing for the nonconsequentialist is easily handled by the most reliable judgment standard.

I should also mention that slavery contracts and religious vows are only the tip of the iceberg of precommitment devices that are not legally enforceable. Consider, for example, the possibility of a contract for permanent marriage without divorce. Or suppose Ron, a young Democrat, is worried that his later self may become a Republican. Ron cannot enter into an enforceable contract to prevent his later self from voting Republican. Or Loni, whose marriage ended in divorce one year ago and for the past year has had the settled values and preferences that would motivate her to voluntarily enter into an enforceable contract to prevent her future self from ever remarrying. There is an endless variety of such contracts, all of them legally unenforceable. None of the standard nonconsequentialist accounts can explain why these contracts should not be enforceable, if entered into voluntarily, because they are backward-looking accounts. The most reliable judgment standard provides a straightforward explanation that distinguishes these kinds of contracts from ordinary contracts.

Addictive Drugs

The question of whether adults should be free to use addictive drugs is a complicated one. The harms from drugs are not solely harms to those who use them. Drug users often neglect their children. Under the influence of certain drugs—for example, alcohol—drivers are much more likely to kill and injure others in auto accidents. On the other hand, even if a ban were justifiable in theory, experience has shown that there are serious practical problems in implementing a ban. I wish to temporarily set aside all such

issues, to focus on the paternalistic issue. Should adults have a right against paternalistic laws banning addictive drugs?

Feinberg would say yes, so long as the decision to use the drugs was voluntary in the sense of expressing the agent's settled values and preferences (1986, 133–134). Those who would justify paternalistic intervention typically point to the addictive characteristics of the drugs, and argue that a person should not be free to compromise her future autonomy in the way an addict's autonomy is compromised.[21] But that can't be right. Some people develop an addiction to coffee, and some people develop what seems like an addiction to exercise. Indeed, if there were a pill that created an addiction to exercise, many people would use it, because they know that they won't get enough exercise if left to their own devices and their future selves would bilaterally endorse their being able to take the pill. Surely, addiction by itself does not justify legal prohibition.

If addiction per se is not always bad, what could be the basis, at least in theory, for legal paternalism to prevent people from experimenting with addictive drugs? The answer is given by the most reliable judgment standard. There might be good reason to think the experimenters' hypothetical future selves would endorse a policy of intervention.

In the previous chapter, I discussed the example of Allen and the recreational drug RD. Suppose it is reasonable to expect that a majority of the future potential users of RD such as Allen would come to unequivocally endorse a drug prohibition, both in the case in which there were such a prohibition and in the case in which there were not. Then the prohibition would satisfy the most reliable judgment standard. It would qualify as soft paternalism.

It is difficult to apply the most reliable judgment standard to actual drug prohibitions, because of the lack of information about the retrospective attitudes of drug users to their drug use. I believe it would be very useful for the government to gather and to disseminate this sort of information on each type of drug, giving the percentage of users who come to regret ever having used it, as well as the percentage who would endorse a ban. It would be necessary to categorize their reasons for regret, so that it was possible to distinguish between the effects of the drug itself and the effects of its being illegal. For example, convicted drug users would surely regret their time in prison. This would be a reason to regret only the drug's being illegal, not to regret the use of the drug itself. If this sort of information were available, it might lead to more informed decisions by potential drug users and perhaps, ultimately, to revisions in the drug laws—or, at the very least, to the elimination of prohibitions on drugs users don't regret using. For example, the prohibition on marijuana use would never survive this kind of scrutiny. Unfortunately, the U.S. government would probably never publish such information, because it would be interpreted as an implicit endorsement of some drugs over others. But such discrimination is essential to defining a right of adults against legal paternalism.

In the absence of the relevant statistical information, my best guess is that although some prohibitions (e.g., on marijuana) would never satisfy the most reliable judgment standard, prohibitions on some others (e.g., methamphetamines) might.

This account differs markedly from backward-looking nonconsequentialist accounts. Because Feinberg's (1986) voluntariness standard looks only at the voluntariness of the decision to use the drugs and ignores the potential opinions of future selves, it implies that people should be free to make choices that, judged by their own future selves, will ruin their lives. Is the autonomy to ruin one's life in the estimation of one's own future self a kind of autonomy that we should respect and celebrate? It may be that the side effects of prohibitions are so serious that no drug prohibitions should be enacted. If so, I would not regard this as an occasion for celebrating autonomy. To me, it would be sad, regrettable, and dismaying admission that we had no effective means to save people from making choices that they themselves will very probably come to judge to have ruined their lives.

Forced Medical Care

Another important issue of legal paternalism is the question of whether it is justifiable to force a lifesaving medical procedure on an adult who rejects it. The classic example of this is forcing a blood transfusion on a normal adult Jehovah's Witness or Christian Scientist against her will. The hypothetical endorsement of intervention standard would require us to defer to the patient's wishes, and this has been the rule enforced by the courts.[22]

However, at least in theory, I don't believe that the current self's judgment should be decisive. At least hypothetically, we can identify evidence that would qualify such intervention as soft legal paternalism under the most reliable judgment standard. Suppose there had been many cases in which normal adult Jehovah's Witnesses or Christian Scientists had been given lifesaving blood transfusions against their will. It would be useful to know how they evaluated their lives after the transfusion. Consider the two most extreme outcomes. In the first, those who received the transfusion all felt their bodies had been polluted. They became despondent and lost their desire to go on living. The most reliable judgment standard would not be satisfied in such a case, and to continue to force them to undergo such transfusions would be hard legal paternalism.

In the second extreme outcome, suppose those who were forced to undergo such transfusions went on to lead happy lives they themselves regarded as worthwhile. The case would be complicated if, nonetheless, they still insisted it would have been better for them if they had never received the transfusion. Suppose they do not. Suppose they come to judge that it was good for them to have been forced to receive the transfusion.[23] Then this would resemble the example of Lee the soccer player, discussed in the previous chapter.

See figure 12.4. The most reliable judgment standard would be satisfied and forced blood transfusions would qualify as soft legal paternalism.

Paternalism and Persons with Disabilities

In the first volume, I presented the history of the development of human rights as a history of overcoming paternalistically justified oppression. The latest chapter in that history is the emergence of human rights for persons with disabilities, represented in the United States by the Americans with Disabilities Act and internationally with the U.N. Convention on the Rights of Persons with Disabilities (2009). Legally respected human rights for those with disabilities have the potential to make a great contribution to the equitable promotion of life prospects.

However, it is important for me to say something about persons with disabilities, because throughout I have characterized the human rights as the rights that should be guaranteed to normal human adults. What do I mean by *normal*? My use of *normal* is not meant to exclude those with disabilities. It is rather meant to emphasize that normal cognitive, emotional, and behavioral development is sufficient for achieving autonomy in my consequentialist sense of the term—that is to have good judgment and self-determination. It is not necessary to be normal in this sense in order to have good judgment and self-determination. Because of the way that historically the epithet *abnormal* has been used by majorities to exclude those who are not "like us," it is important for me to emphasize that I intend my use of *normal* to be inclusionary rather than exclusionary. On my account, everyone should be presumed to be capable of autonomy (in my consequentialist sense) and should be assured the rights necessary to develop it. Only when, in spite of our best efforts to develop it, the evidence is conclusive that a person lacks good judgment or self-determination should he be categorized as lacking autonomy. Those who lack autonomy do not lack all rights, but their rights are different from the human rights of autonomous adults.[24] For example, legal paternalism, such as the appointment of a legal guardian, is appropriate toward those who are not autonomous.

Why *Rights* against Legal Paternalism?

In a democracy, a paternalistic law can be enacted by a majority. This is as it should be when a majority endorses the law because of the benefits of paternalistic intervention in their own case, and not because the majority thinks the law will be good for a minority that does not and will not come to endorse it themselves. In the former case, a minority may be bound by a law it does not endorse, due to the majority spillover effect. Such laws do not violate any right of the minority.

In the latter case, a majority uses the legal system to give effect to what it believes is good for *other people*. Because in a democracy majorities have the power to enact laws, so long as they are constitutional, only a constitutional right to protection against legal paternalism can effectively protect minorities against a majority's enactment of legal paternalism toward the minority. For this reason, there is no other way to protect a minority against legal paternalism than by incorporating a robust right against legal paternalism into the constitution.

To the credit of the U.S. Supreme Court, it has identified a constitutional right to a sphere of autonomy free of some legal paternalism. The most reliable judgment standard is an appropriate standard for the Supreme Court to use to define this right against legal paternalism, because it is not the Court's role to make judgments about whether the negative side effects of such a law outweigh its benefits. That is for the legislature to determine. The role of the Supreme Court is a filtering one: to invalidate legislation in which a majority improperly imposes its judgment about what is good for a minority on that minority. The most reliable judgment standard provides a standard for that filtering role that, for the reasons discussed in this chapter, is superior to the alternative nonconsequentialist standards. On the basis of the most reliable judgment standard, it is possible to define a human right against legal paternalism for normal adults to include the following:

1. A right to religious freedom
2. A right to sexual freedom
3. A right to reproductive freedom
4. A right to refuse medical treatment, including a right to refuse extraordinary care and to be removed from life support
5. A right to marry that includes same-sex marriage
6. A right to suicide and assisted suicide in certain end-of-life situations

Two items that are not on the list are drug prohibitions and suicide when contemplated outside of end-of-life situations. Although the main principle creates a presumption against drug prohibitions, if they are designed to be humane, it does not rule them out entirely. However, it creates a presumption *in favor of* laws to prevent suicide, if they do so humanely.

Privacy Rights

The liberty right against legal paternalism was originally conceptualized as a privacy right. Once it is recognized to be a liberty right, not a privacy right, we can ask about privacy rights themselves: Are there any other privacy rights that should be human rights?

There are two potential candidates: (1) a right to a private space, protected from physical intrusion and certain other kinds of access (e.g., wiretapping)

and (2) a right to informational privacy, that certain kinds of content be protected from being revealed (e.g., medical records).

Griffin thinks that only the second kind of right, a content-based right to informational privacy is a genuine human right (2008, 235). That is because his conception of a human rights is based on the conditions for being a normative agent, and he does not think that violations of a private space are serious enough to compromise one's status as a normative agent. I have proposed a different criterion for a human right. They are the rights that the main principle endorses as universal, inalienable robust legal protections against government or majority tyranny. Let's consider three elements individually.[25]

(1) *Robustness*. A robust right is one that cannot by overruled by a government official or by a simple majority. Consider first the right to a private space. This right has two aspects. The first is the requirement of a warrant to search one's private space. Historically, this has been an important protection against government tyranny, so it easily qualifies as a robust right. The other aspect is the potential for using technology to intercept private communications or to access private activity. Such technology did not exist at the time the U.S. Constitution was drafted, but the courts have extended the right against unreasonable searches to cover this kind of privacy invasion, because it has the same potential for government tyranny as physical invasion of one's private space. So I conclude that the main principle would favor a robust right of this kind.

What about informational privacy? For example, should governments be able to collect DNA profiles on all of their citizens without their consent? In theory, it would seem there would be lots of benefits to being able to do so. However, concern about potential abuses seems to warrant a right against the government and against a majority in the legislature. I am less certain about this case, but, on balance, it seems to me that the main principle would also favor a robust right of this kind.

(2) *Universality*. It is easy to see why a right to a private space would be universal, at least in modern society, because protection against warrantless searches would be an important protection against government tyranny in any modern society. It may be harder to see that there is a basis for a universal right to informational privacy. However, if it was ever thought that some cultures don't have a conception of informational privacy, the AIDS epidemic has shown that to be false. In every culture, those who are HIV positive try to keep that information private. It is clear that they have an interest in doing so. There may be cases in which other rights take priority over the right to informational privacy, but that is just to say that the right is not absolute, but robust.

(3) *Inalienability*. Of course, people should be free to allow others into their private space or to share private information with others. Should they be free to contract away their privacy rights altogether? This seems to threaten the kind of abuse that makes inalienability a solution to a CAP. It is easy to

imagine that an employer might require that an employee give up informational privacy as a condition of employment. Or an employer might provide dormitories for workers and then claim that they had no private space, because the employer owned the dormitories. In either case, it would be important for the relevant privacy right to be inalienable. So I conclude that the main principle would endorse both kinds of privacy rights as universal, robust, inalienable rights—that is, human rights.

FOURTEEN

Clarifications and Responses to Objections

The Moral Significance of Borders

In this book, I have deepened the consequentialist account of basic human rights from the first volume and extended it to a longer list of basic and nonbasic human rights. My account of human rights qualifies as an *institutional* account (e.g., Nagel 2005), because I conceive of them as rights that all governments should guarantee to everyone everywhere. Because my account depends on governments to guarantee the rights, it is institutional.

Some advocates of human rights believe that securing human rights is the responsibility not only of governments, but also of individuals. No such account has adequately addressed the problem of nonideal theory discussed by Murphy (2000). It is quite plausible that I might have an obligation to contribute my fair share to provide medication for those who are HIV+ in Africa. It is not plausible that if no one else contributes their fair share then I am responsible for all providing medication for as many of those who are HIV+ in Africa as I can support on my salary.

Because my account of human rights grounds them in moral reciprocity, it may seem that my account has no implications for human rights across borders. It is true that if there were two isolated societies with no potential for mutually beneficial interaction, my account would imply that there were no moral reciprocity relations between them and thus the main principle would not apply to relations between them. One society might well have humanitarian duties toward the other, but those would not be covered by the main principle.

However, the world we inhabit is not like this. In the world we inhabit there is a vast web of economic and social relations between the members of different states. Those relations are governed by coercive enforcement of international law and custom—including, for example, laws of property and contract. The main principle applies to those reciprocity relations, but it applies to relations between individuals, not states (cf. Blake forthcoming). Even if, contrary to fact, it were true that international trade equitably divided the gains from trade among governments, if those governments did not translate the gains into policies that equitably promoted the life prospects of their citizens, the main principle would favor an alternative arrangement that did so.

Thus, for example, the current system of international law permits a dictator to dispose of his country's natural resources or incur national debts even if the proceeds do not promote the life prospects of his citizens. Although implementation problems would be immense in the current state of the world, it is easy to imagine that sometime in the future international law might be amended to limit the right to dispose of natural resources or to incur national debts to democratically elected governments that guarantee the basic human rights (cf. Pogge 2002). If the implementation costs were not too great, such a change would almost surely be endorsed by the main principle. By limiting the benefits of being a dictator, such a change would reduce the incentive to be one.

There is a great opportunity for new institutions and new legal arrangements that have the effect of equitably promoting life prospects globally. It is a mistake to think that we must wait for state actors to take the initiative. State actors are often so focused on national self-interest that they cannot reach agreements to promote equity, or if they do, they are loathe to comply with them. The glacially slow response to climate change, consisting mostly of unfulfilled promises, is a prominent example.

However, the movement to promote life prospects globally is not waiting for state actors. Like almost all movements for moral progress in history, it is largely bottom-up. In chapter 9, I pointed to the bottom-up change in consumers that has made it necessary for corporations to compete not only on product quality and price, but also on fair trade and environmental protection. This is one of the most powerful forces for human rights in the world, and it pays no attention to national borders.

In addition, there are many trans-national initiatives from nongovernmental organizations. Consider, for example, the work of the Gates Foundation and other private foundations working on global health. They are taking the lead in establishing the recognition of a global right to health care. Or consider Tostan, a nongovernmental organization that began in Senegal. Tostan has used education in human rights as a model for empowering women. As a result of Tostan's work, women in Senegal are using their power to improve sanitation and end female genital cutting and child marriage. Tostan has expanded with programs in Burkina Faso, Djibouti, Gambia, Guinea, Guinea Bissau, Mali, Mauritania, Somalia, and Sudan. We can now foresee the day when at least the most severe forms of female genital cutting have been eliminated from the earth. This bottom-up movement is yet another illustration of how the human rights movement has made states and national borders less important.

Of course, global exploitation won't end until there are institutions to detect it and sanction it. So, ultimately, the bottom-up global movements will need to establish some institutions to secure and protect their achievements. This need not lead to a global government, but will at least require global human rights enforcement agencies, perhaps modeled on the International Criminal Court.

Nagel (2005) is right that an institutionalist about human rights must acknowledge that, where governments do not secure human rights, the international institutions to secure them by and large do not exist yet. I would simply add to Nagel's account that the idea of human rights can play an important role in the bottom-up processes of transformation that improve life prospects and are themselves the best hope of bringing into existence the institutions that will legally guarantee human rights.

Is My Account Really Consequentialist?

I have claimed that my account is a consequentialist account of the content of human rights norms. But is it really a version of consequentialism, or is it rather a veiled form of nonconsequentialism? There are four reasons for thinking that it is not really consequentialist: First, my account is not consequentialist about all of morality, only about changes in moral traditions that have passed the consequentialist threshold; second, even for moral traditions that have passed the consequentialist threshold, it only applies to changes in primary ground-level moral practices, not to secondary moral practices (having to do with enforcement); third, I only claim that satisfying the main principle is a sufficient condition for a change's being a moral improvement, not a necessary and sufficient condition; and fourth, because my account is based on moral reciprocity relations, it depends on a distinction between persons and non-persons and a distinction between cooperators and noncooperators. For both distinctions, my account depends on an independent account of moral responsibility. I offer no account of moral responsibility.

This is a reminder that my account is not an account of all of morality. It is not an account of personhood, or the grounds for punishment (moral responsibility), or the proportionality and other constraints on permissible punishment or enforcement (secondary norms and judgments). It is not even an account of moral obligation, because it provides only a sufficient condition for a change in the status quo moral practices to be an improvement. At most, it explains when such a change would be morally permissible, but it does not provide any way of determining when such a change is morally obligatory.

But even if it is not a consequentialist theory of morality or of punishment or of all moral improvement, there is a sense in which it is a consequentialist theory of something. It is an objective theory of improvements in the primary moral practices of a social group that has passed the consequentialist threshold, because it makes the determination of whether or not they are improvements dependent on the satisfaction of a multiple-time-slice end state principle—that is, a function of how well they (evaluated as a substantive practice and a practice of implementation) equitably promote the life prospects of cooperators (compliers and nonresponsible noncompliers, when the responsible noncomplier exclusion applies) in comparison with

the relevant alternatives. And it is the principle that explains why it would be a moral improvement for all governments everywhere to adopt the fourteen human rights norms on my list. That is all I mean when I say that the account of the content of human rights norms is consequentialist.

Moral Sensitivity

The main principle is an objective consequentialist principle of moral improvement. It classifies changes to moral practices as improvements based solely on an evaluation of the substantive practice and the implementation practice. Thus, the main principle can evaluate a change to be an improvement even if the change happens by accident.

It is probably true that early in human development, random changes in moral practices were favored or disfavored by biological selection and "improvements" were those that were advantageous in that process of selection. However, at some stage of development, something new appears. At least some people are able to (fallibly) recognize that exceptions to their moral norms are sometimes warranted. Because traditional moral norms are generally regarded as coming from an infallible source, their own self-understanding of this ability is generally not that they are finding exceptions to norms, but rather that they are more carefully interpreting the existing norms. In any case, these exceptions or reinterpretations are evidence of sensitivity to the consequentialist considerations stated explicitly in the main principle. It is at this point that a moral tradition crosses the consequentialist threshold. From this point on, changes in the moral practices are no longer random or purely accidental. The moral sensitivity that makes it possible to recognize exceptions will never be infallible, but, over time, it provides the basis for a nonrandom process of objective moral improvement.

There is no one-to-one correspondence between changes in moral practices that result from this moral sensitivity and objective improvements. A moral change based on this kind of moral sensitivity could have unforeseen consequences that make it a moral mistake, as was Marx's proposal to abolish private property. On the other hand, change made for morally repugnant reasons could turn out to be an objective improvement, as was the Catholic Church's permitting Christians to pay interest on money borrowed from Jews, allowed on the grounds that the Jews were already doomed to go to hell.

In the first volume, I illustrated this fallible ability to recognize exceptions to moral norms with the example of Bartolomé de las Casas. By the end of his life, las Casas had come to the conclusion that, not only was it wrong for the Spanish to use force to convert the American natives to Christianity, but that it had been a mistake to seek their voluntary conversion also. This conclusion conflicted with one of the main norms of his own religion. He did not come to this conclusion by reasoning from other norms or principles.

He came to the conclusion by way of empathic understanding of the natives that enabled him to appreciate how devastating the effects of the conversion to Christianity had been on their ways of life.

This process is dramatized in Mark Twain's [1884] fiction, in the story of Huck Finn's friendship with the slave Jim and how that friendship led him to make an exception to the moral norms that structured his life—that slaves were property and that helping a slave to escape was stealing. Bennett (1974) draws our attention to how this simple story undermines the Kantian picture of morality based on principles or norms or conscience, because Huck's norms and his conscience told him he should turn Jim in. And just to make the Kantian reference explicit, Twain included in the story two white men searching for runaway slaves who ask Huck if he has seen any. Everyone familiar with Kant's [1799] moral theory knows that he explicitly drew the conclusion that it is always wrong to lie, even when such bad consequences can be foreseen to follow from telling the truth. Huck never for a minute has the thought that it is permissible for him to lie. He just can't get the words out, he can't tell them the truth. That is what moral sensitivity feels like. With one simple story, Twain cast more doubt on the Kantian conception of morality than any philosophical argument could. He also showed why the main principle applies to practices that produce moral sensitivity, rather than to norms or principles per se.

Because of the important role of empathy in the true story of las Casas and the fictional story of Huck Finn, it may seem that the Humean account of morality as merely an expression of feelings is correct. In the first volume, I explained why I think that is a mistake. Hume thought that if feelings were involved in moral judgment, then it could not be a judgment of anything objective. But that was a mistake. The kind of moral sensitivity exhibited by las Casas and by Huck is mediated by feelings, especially feelings of empathy, but it is at least possible that it is itself a (fallible) sensitivity to something objective, the main principle.

One of the great successes of human cultural development is the fact that this kind of moral sensitivity can be developed in any human culture. There is both direct and indirect evidence of this sensitivity. The direct evidence is that people in any culture can feel the tension in their own ground-level moral norms that this kind of sensitivity generates. For example, women in almost any patriarchal culture can feel that there is something unfair about their position, even if, like Huck, they cannot articulate what it is. The indirect evidence is the elaborate rationalizations that cultures construct for their own cultural practices in order to be able to allay the feelings generated by this kind of moral sensitivity. In the previous volume, I discussed how the vehemence with which cultures suppress any questions about the justification of their cultural practices is itself evidence that this kind of rationalization is occurring. Every culture rationalizes its own practices and attempts to suppress questions about their justification. The recognition of human rights is a transformative development in the process of moral improvement because

it provides the protection necessary for people to question their moral norms without risking severe retribution, and thus greatly facilitates the process of reforming cultural practices.

Before a culture attains the kind of moral sensitivity evidenced by the adoption of some form of the Golden Rule, improvements in its moral code are random and there is even a sense in which it is not a genuine moral code. Once a culture attains some minimum level of this kind of moral sensitivity, then the culture itself is the engine of its moral improvement and, because the improvement is due to the moral sensitivity of the members of the culture, the members of the culture are genuine moral beings and their code is a genuine moral code, rather than merely a culturally transmitted set of attitudes, responses, and behavioral constraints.

How much moral sensitivity is required for a culture to have a true moral code? I use the Golden Rule as a positive test, because a society that advocates some version of the Golden Rule has achieved at least an implicit awareness of the considerations made explicit in the main principle.

My Own Theoretical Inertia and My Own Fudge Factors

In chapter 2, I suggested that Nozick was a victim of theoretical inertia in his treatment of the example of Marie the medical researcher. In earlier chapters, I suggested that Mill and Rawls employed fudge factors in their theories. One of the reasons I did this was to draw attention to my own theoretical inertia and my own fudge factors. Sadly, I am not in a good position to spot them. I depend on your help.

Let me start with theoretical inertia. One of the great advantages that I have as the author of this book is that I get to choose the examples that I discuss. It would be dishonest of me to hold back examples that I knew were problems for my view. But theoretical inertia is more subtle than that. It is not that I think of counterexamples to my view and try to keep them from seeing the light of day. I just tend to think of examples that support my view more readily than examples that do not. When I come upon an example that seems to me to support my view, it makes a real impression on me. Examples that cast doubt on my view will be harder for me to recognize. These are familiar cognitive biases. As a result, it is up to you to step back and think of examples that cast doubt on my view. You can't rely on me to do a good job of that.

It is a little embarrassing to think that I might be unconsciously favoring my own theory, but it is a reminder that we depend on the free give-and-take of opinion to help winnow the good from the bad. I make my contribution to that process not to end it, but to nudge it in a slightly different direction. As I said in chapter 7, almost everything in this book is subject to reasonable disagreement.

What about fudge factors? Well, it seems that I have two pretty big ones, well-being and equity. Because I don't have a definition for either of them, I

have the flexibility to respond to some potential counterexamples by emphasizing well-being and to others by emphasizing equity. Is this fatal to the theory?

It might be if I just used the theory to draw conclusions that everyone already agreed with antecedently. Would there be any point to such a theory? There might be. Well-being and equity are only two variables. If it were possible to explain all or almost all moral improvements with only those two variables, that would be a surprising discovery. However, I have not just used those variables to explain what was already explained nonconsequentially. I have used them to challenge nonconsequentialist explanations in almost every chapter of this book. The challenges include: the rationale for a criminal justice system known to punish the innocent and the potential for criminal and civil strict liability, the rationale for the apparent retroactivity of some decisions in civil and criminal appeals, the rationale for a right to freedom of expression that covers intolerant subversive advocacy, the rationale for the unconscionability doctrine and for the various other doctrines that are used to set aside contracts voluntarily entered into, the rationale for favoring *ex post* over *ex ante* consent on win-win contracts, the potential rationale for replacing democratic rights with election by deliberative poll, the rationale for exceptions to Feinberg's nonconsequentialist standard for weak paternalism, and the rationale for why human rights should be inalienable, to mention only some of the challenges.

Do I really believe that there is a definition of well-being and a formula for equity? Not a definition in the semantic sense. In that sense almost none of our words have definitions. But I do believe that there are important truths about well-being to be discovered and either a formula or something like it for equity at the meta-level.

Nagel (1991) despairs of any such formula, because he imagines it being applied in a kind of Scanlonian original position, where, without a veil of ignorance, everyone comes together and discusses what trade-offs of well-being for equity there are that no one could reasonably reject. This discussion would bog down quickly, because these people all have their own life projects based in part on their current position. So to successfully promote equity, the discussion would have to persuade them to give up their life projects, which they quite reasonably might be unwilling to do.

This is just another example of the mistake involved in thinking that the main principle is a ground-level moral principle. I believe that Nagel is correct that there is no ground-level moral principle for trading off well-being and equity that would gain unanimous consent in a Scanlonian original position. One reason for this is that there is a powerful anchoring phenomenon in human thinking. When I think about changes to the status quo, I think about whether they would improve or worsen my current position. There is a limit to how far even a reasonable person will be willing to worsen her position from the status quo.

Now consider a different way of approaching the problem. Each generation gets together in a Scanlonian original position and the better off are

asked to accept a 1% increase in the highest marginal tax rates to improve the position of the least well off. If there are significant inequalities, this seems like a relatively small departure from the status quo, and it will certainly seem unreasonable if the better off won't agree to it. It is hard to imagine that anyone among the better off would have his life projects significantly affected by a 1% drop in his income. However, over a large number of generations, this process can reach whatever the optimum tax rate would be, as determined by the equity formula in the main principle.

Nagel considers this kind of gradual improvement in equity, which he calls a "benign slippery slope" (1991, 90), but is pessimistic that it could be brought about by majoritarian democracy. It is important to notice that this is a different kind of problem than the problem of reasonable rejection in a Scanlonian original position. If the "benign slippery slope" can solve that theoretical problem, then it is much more plausible that there is a formula for equity. Of course, it would be unfortunate if there were a formula for equity but no political process could realize it. On the politics, I am more optimistic than Nagel, at least over the longer term.

If a formula for equity were ever discovered, I would expect it to have little, if any, effect on our ground-level moral and legal norms. Certainly, they would not be modified to incorporate the formula. As I explained in chapter 5, the main principle is not a ground-level principle to be applied in our moral reasoning. It is a meta-level principle. In our ground-level moral practices, we depend on developing a sensitivity to the requirements of promoting well-being and equity and of trading off one for the other and of acting on that basis. A formula for equity or for trading off equity against well-being would not be of much help, because the main principle would not endorse setting up a government office to apply it. This is the moral of the example of the bureau of equity in chapter 3.

Thus, in our ground-level moral reasoning, I doubt that we will ever have anything more than rough rules of thumb for balancing gains in well-being against equity of distribution. But if there were no meta-principle that our practice was sensitive to, then our rules of thumb and our ability to make exceptions to them would be arbitrary. It is not necessary to believe that there is a complete ordering of systems of social practices based on the extent to which they equitably promote well-being. Almost everyone would agree that there is at least a partial ordering. It is clear that familiar slavery systems, caste systems, and patriarchal systems are lower on the scale than some nonslave, noncaste, nonpatriarchal alternatives. How complete is the ordering? Complete enough for the historical process of making improvements to go on for a long time.

Let me conclude with a reminder that it is almost inevitable that my view has developed its own theoretical inertia and probable that I have sometimes used well-being and equity as fudge factors. I depend on you to be more sensitive to those problems than I can be.

Another Fudge Factor: A Circularity Problem

There is another potential fudge factor in my account, due to the following circularity problem: My theory purports to be a theory of moral improvement. My test of the theory is how well it fits my beliefs about which moral changes have been moral improvements. Is it any surprise that there is a good fit?

This circularity problem is exacerbated by the fact that, since we all have to agree that we have moral blind spots, it would actually discredit my theory if it agreed with all of our current opinions on which changes would be moral improvements. So if anyone finds an implication of my theory that conflicts with current opinions on how to improve our moral practices, I can just play the moral blind spot card and save the theory.

This would be a serious problem if, for example, I had to give equal weight to everyone's opinions about moral improvement. If the Taliban's opinion that it is a moral improvement to deny women an education and to keep them almost imprisoned in the home was as valid as mine, my theory would never get off the ground. In the first volume, I explained why all moral views are not equally valid, and that explanation applies directly here.

However, even if all moral views are not equally valid, there are lots of nonconsequentialist views that are as valid as mine. Recall that in chapter 1, I acknowledged that there is a presumption in favor of the nonconsequentialist views, because they more closely fit our ground-level moral reasoning. For my theory to be successful, it is not enough that it fits my beliefs about which moral changes have been improvements. I have to find examples that will put pressure on the nonconsequentialist views. That is why I have focused most of my discussion on examples that I thought most nonconsequentialists would at least feel the force of. And also why I focused special attention on the value of choice and on a right against paternalism. Those are the areas where the nonconsequentialist seems be on strongest ground. If I can raise doubts about the nonconsequentialist explanations there, then I am clearly not just reinforcing my own consequentialist preconceptions. Ultimately, the test of whether I have avoided the circularity objection is whether nonconsequentialists think that I have raised challenges that they need to respond to.

Is My Indirect Consequentialism Incoherent?

In chapter 5 I argued that the main principle does not endorse our using it as a ground-level moral principle, because human beings make big mistakes when they try to apply it. Then in the rest of the book, I have been applying the main principle. If I am able to apply it without making mistakes, why can't others?

To be a consequentialist, I have to be able to apply the main principle to explain historical improvements in law and morality. I don't have to be able to predict the effects of legislation on the equitable promotion of life prospects. It is much easier to use the main principle in retrospective explanation than in the kind of prospective prediction required for legislation.

Another way to put the point is the way that I put it in the first volume. Of course, I should be free to offer my account of the main principle in the free give-and-take of opinion. But suppose I woke up tomorrow and I had been made dictator of the world. Tempting though it might be for me to use those powers to try to implement the main principle, the main principle would not endorse my doing so. Instead, it would endorse my acting so as to establish the institutions necessary for a transition to a democracy with constitutional protections of human rights.

More Doubts about the Claim of First-Person Authority

Is the claim of first person authority really true? In the first volume I considered some reasons for thinking it is true and responded to objections. The most common misunderstanding of the claim is to think that it implies that people are good at predicting the future. If it had this implication, it would be clearly false. Indeed, one of the grounds of human rights is that no one is very good at predicting the future, not even the experts. When Tetlock (2005) did a careful study of the reliability of predictions by political experts, he found that they were little better than chance. This is a further reason that the main principle does not endorse rule by political experts. If no one's predictions about the future are very good, then the only way to reliably improve social institutions is to make piecemeal changes in them and then get feedback from those who are affected by them on whether the changes have improved or diminished their life prospects. For the same reason that the most reliable judgment standard favors people's retrospective judgments about the benefits of a paternalistic law, the main principle favors a political system that is responsive to people's retrospective judgments on how a law has affected them. Of course, there is always the potential for unanticipated future effects and this problem, which in chapter 10 I called the *time lag problem*, is a problem for democracy, as for any other form of government. But a rights-respecting democracy at least has the advantage of the process of the free give-and-take-of-opinion to identify such problems.

The First Problem of Contingency: Trade-Offs

One of the most compelling objections to consequentialist accounts of rights is that they allow trade-offs of rights violations to promote well-being. Since the guiding idea of this entire book is that we learn about the main principle

by paying attention to the process of incorporating exceptions into our moral and legal norms, I have to agree that all ground-level moral and legal norms potentially have exceptions. Even nonconsequentialists such as the early Nozick (1974, 30n) allow for some exceptions. Also, any effective criminal justice system is going to convict a substantial number of innocent defendants. Everyone, consequentialist and nonconsequentialist alike, needs some way of dealing with trade-offs.

Direct utilitarianism has the most morally disturbing implications for trade-offs. My version of consequentialism avoids the morally unacceptable implications of direct utilitarianism for two reasons. First, it is an indirect consequentialist view. The main principle applies to practices, not to individual decisions of government officials. It endorses ground-level moral and legal principles that establish robust rights that take precedence over the judgments of government officials and the judgments of simple majorities that infringing a right would promote well-being.

Second, unlike utilitarianism, the main principle pays attention to the distribution of well-being. Those who have their human rights violated are typically among the worst off. Since the main principle gives priority to the life prospects of the worst off, it will give extra weight to protecting human rights.

Nonetheless, it is true that, because the main principle endorses human rights on the basis of their role in a practice that equitably promotes life prospects, it could potentially endorse practices that infringe on rights in a way that does not have too great an effect on well-being. One somewhat surprising example of this very possibility is the public figures exception to privacy rights, which has developed as a judicial doctrine in the United States. It is not surprising that there would be an exception to privacy rights of government officials or to people who just happen to be involved in newsworthy events. Clearly the public has an important interest in information that requires limiting privacy rights in these cases. The surprising part of the public figures exception is that it applies to the private lives of celebrities. The rationale given by the courts for this exception is that by placing themselves in the public eye, celebrities implicitly consent to giving up their rights to the nondisclosure of private facts about their lives, with some exceptions (American Law Institute 1977, section 652D, comments e, g, and h). This rationale would be ludicrous if given in any other context. No one thinks that by going for walks at night people implicitly give up their rights not to be mugged.

The only plausible rationale for such an exception is consequentialist. Although loss of privacy for celebrities is an annoyance for them, the fact that they are among the most well off means that this exception does not raise any problem of equity. In fact, in a strange way, because it raises the life prospects of so many of the less well off who enjoy reading about the private lives of celebrities and seeing unposed photos of them, it promotes equity.

So if the claim is that consequentialism permits trade-offs of rights against well-being in some cases, the reply is that any reasonable view must do so. The challenge for both consequentialist and nonconsequentialist accounts is to explain the difference between the trade-offs that are permissible and those that are not.

The Problem of Divided Reason

Anderson objects to any indirect consequentialist account of moral and other values that it "fails to provide us with a coherent basis for self-understanding and requires disturbing divisions among different aspects of the self" (1993, 43). She raises this objection because she thinks that the attempt to explain moral and other values in terms of their contribution to well-being deprives them of the value that we take them to have in our ground-level judgments.

Suppose that I have the ground-level moral belief that I ought to help my friend when he needs help. Anderson would say that my indirect consequentialist account does not vindicate that ground-level belief, because it does not imply that I *really* ought to help my friend, rather it explains why it is good or appropriate for me to *believe* that I ought to help my friend. But that means that the indirect consequentialist account *undermines* the ground-level belief, by implying that, strictly speaking, it is false. Indirect consequentialism makes morality an illusion. A useful illusion perhaps, but an illusion nonetheless. It requires us to have a divided consciousness. In one part of consciousness we believe that we ought to help our friends when they need help, while in the other part we have to admit that the reason that we should hold that moral belief is that when people generally hold such beliefs, the consequences are good. According to Anderson, this kind of divided consciousness is "repugnant to common sense" (1993, 43).

Anderson is claiming that the indirect consequentialist introduces a division into reason, in this case, a division in practical reason. I agree that it does. Therefore, if there were an *a priori* argument that there can be no divisions in reason, my position would be refuted. However, there is no such *a priori* argument. On the contrary, when we look closely at reason, we find more than one division. So we need to consider whether it could be reasonable to have a division in practical reason.

My consideration of this question has three parts. First, I consider what we are to make of explanations that common sense finds repugnant. Second, I discuss an analogy between the division in practical reason that Anderson finds "repugnant" and a similar division in theoretical reason. The analogy is meant to make it plausible that mature intellectual development requires reconciling oneself to some divisions in reason. The third part of my reply is to consider a different division in reason, one that all accounts, including Anderson's, have to acknowledge. I will suggest that my indirect consequentialist account of moral improvement has the potential to bridge this third division.

Some True Explanations are Repugnant to Common Sense

The first part of my response to Anderson is to simply point out that common sense has in the past found true explanations repugnant—for example, the Galileo's explanation of the motions of the planets. Why should our understanding of ourselves as moral beings and as valuers not also lead us to discoveries that common sense finds repugnant? In her own account of value, Anderson does a good job of describing the phenomenology of ground-level value judgments and of ground-level conflicts of values, but she provides no explanation of how such conflicts should be resolved and no explanation of how our value judgments can be improved. Consider the example of the value of friendship. Although George may now be my friend, if at some point George's needs become so great that they threaten to take over my life, it may be that I will need to restore some balance to my life by being less of a friend. Of course, Anderson would allow that the demands of friendship can be outweighed by other values. What is the explanation of this kind of balancing and how do we do it? From within common sense, which is the perspective that Anderson occupies, about all that we can say is that when we are properly attuned to value, we just know how to do it. This was Aristotle's account of virtue as practical wisdom. I think both of them get the phenomenology right, but the phenomenology does not settle the question of whether there is an explanation at the meta-level for the kind of judgment exercised by the person of practical wisdom.

A Division in Theoretical Reason

The second part of my response to Anderson concerns an analogous division in theoretical reason, one that I discussed in chapter 7. For a simple example, consider the preface paradox. I could have written a preface to this book in which I said: I believe everything that I have written in this book, but I also believe that much of it is false.

Clearly, my preface would make the entire book inconsistent. Consistency is often thought to be a requirement for a coherent set of beliefs and thus a constraint on theoretical reason. But it seems to me that it would be irrational of me *not* to acknowledge that some of my beliefs (including some of the assertions in this book) are false. To do so, requires that I introduce some sort of division into theoretical reason.

The same kind of division in theoretical reason was necessitated by my reflections on Mill's epistemology in chapter 7. There I asked how we should think of our own opinions, when we view ourselves as part of the process of free give-and-take of opinion. I suggested that we needed a more complex attitude toward ourselves than the simple attitude of: If I believe it, it is true. If we think of ourselves as part of a process that tends toward the truth over time, then we will regard our own opinions as to some extent provisional, subject to improvement. This attitude conflicts with the common sense view

that to believe something is to believe it true. My suggestion is that if common sense insists on this simple-minded approach to belief, then common sense needs to be modified to allow for more complexity. This division in theoretical reason seems to me not a sign of undesirable division of consciousness, but rather of a mature intellectual development. My suggestion is that mature intellectual development also reveals to us a division in practical reason and that common sense can adjust to it, also.

Bridging the Division between the Moral World and the Natural World

In my third response to Anderson, I consider a different division in reason—one between practical reason and theoretical reason. This division is a troubling one. I believe that indirect consequentialism is our best hope for bridging this division and unifying our self-understanding. I begin with a part historical, part mythical explanation of the source of the division.

There was a time in human moral development when morality seemed a simple matter of following moral norms. Moral beings thought of themselves as bound by those norms because of they were autonomous beings and moral norms were the laws that autonomous beings give themselves. The content of the norms was given by the limits on behavior that were necessary to adequately respond to the immeasurable value of each autonomous being. This state of moral integration was called the *Garden of Eden*.

The Garden of Eden was not a place, it was a state of mind. Humans left the Garden of Eden when they first recognized that there were exceptions to their moral norms. For the first time, they had acquired genuine knowledge of good and evil, but because it was a departure from their earlier uncomplicated state, they regarded it as a loss. They no longer had the feeling that there was a simple formula for how to treat beings of immeasurable worth like themselves. But they still shared in a state of mind in which autonomous beings had immeasurable value. That value could be directly perceived whenever one was in the presence of an autonomous being.

The second great fall from grace was the development of science. Science provided an alternative way of understanding the world that did not involve autonomous beings and moral values. According to physics, everything, including people, is just elementary particles behaving in accordance with universal deterministic or probabilistic laws. This introduced a fundamental division into human consciousness. We could look at the world as made up of elementary constituents governed by the laws of physics (the *natural world*) or we could look at the world as containing autonomous beings of immeasurable value (the *moral world*). The scientific advance that completely severed the two worlds was Darwin's theory of evolution. Then, for the first time, it seemed there could be a fully scientific explanation of the emergence of human beings that did not imbue them with any value and did not need to posit any objective moral norms at all. Of course, there would be

scientific explanations of human moral practices, but the explanations would be debunking explanations, because they would explain why morality was an illusion.

And so, at least since Darwin, the most common way of avoiding a divided consciousness has been to dismiss the objectivity of the moral world altogether. Anyone who ascribed objectivity to the moral world would have to either deny the objectivity of the scientific world or pay the price of a divided consciousness.

There have been many attempts to heal the breach between the two worlds and somehow bring them together. There are two ways to do this. One would be to deny that either of the worlds is objective. The moral world would lose its second-class status if the natural world weren't objective either. This strategy was doomed to fail. Not everything could be a social construction. There had to be something to do the constructing.

The other alternative is to find a way to make sense of the idea of an objective moral world. Nagel (1986) has given us the best way of thinking about this question, in terms of our ability to step back from our personal point of view to think about how things are from an impersonal, objective point of view. When he first employed his thought experiment, he found some objective values and disvalues, the pleasures and pains that the utilitarians had identified. But the most important things that he saw from that point of view were reasons. Reasons seemed to have no place in the natural world. And so, even though Nagel could not claim to have bridged the two worlds or to have integrated them, he at least tried to reestablish the moral world as objective in its own right.

And that is where we are today. Those of us who take the moral world seriously have a divided consciousness. Indirect consequentialism is not the cause of this divided consciousness. On the contrary, it is perhaps the most promising avenue for trying to heal the division. Let me explain why.

Nothing short of a soul will underwrite the infinite value of autonomous beings, and there are no souls in the natural world. But there are conscious beings, some of whom are capable of reasoning and exchanging reasons with others. It is too reductive to say that only pleasure and pain have value for these beings. There are many different sources of value. The utilitarians mistakenly thought that if there was value in the world, rationality required maximizing it. In Rawls's famous words, they did not "take seriously the distinction between persons" (1971, 27).

The distinction between persons is an objective distinction. If we allow any kind of objective value in the world, we see that there is a possibility of collective action problems. Collective action problems can be characterized objectively. Individually rational beings, in the sense employed in rational choice theory, would not be able to solve their collective action problems. The main idea of Rawls's theory is that there is a principle of reasonableness that determines a fair division of the benefits and burdens from cooperation in practices that are solutions to collective action problem. When Rawls

initially presented the theory, it looked as though that principle of reasonableness would apply to all rational beings, and thus that his theory would be metaphysically immodest. Rawls (1985) corrected that impression when he gave his theory a political, not metaphysical, interpretation. But that only left open the possibility for someone else to take up. If there is a principle that determines a fair division of the benefits and burdens of cooperative social practices that applies to all rational beings, it will be a principle for equitably promoting well-being, such as the main principle. So accepting the truth of my indirect consequentialist account of moral and legal improvement has the potential to heal at least one division in reason. The division is already there. My account is intended to help heal it.

It is true that the healing will require a change in many people's understanding of the moral world. Our valuing of other autonomous beings will be explained not by their having immeasurable value, but rather by what is owed to them from one valuer to another in our cooperative endeavor to live worthwhile lives. And human rights will be seen not as essential to being human, but rather as providing a framework in which we can live together, conduct experiments in living, and benefit from the experiments of others.

This is a change in self-understanding. Is it a divided self-understanding? Perhaps. When I think of you as having rights, it does not seem to me to in any way to diminish your rights if I think of them as the mutual constraints that enable us to fairly share the benefits and burdens of cooperative social practices—that is, to participate as equals in a system that equitably promotes well-being. But I do acknowledge that if you found people in the Garden of Eden and suggested to them that my indirect consequentialist account was the best way to understand human rights, they would think you were crazy.

Other Values?

On my account, human rights, and indeed morality and law itself, are a framework within which people can make and pursue their own life plans. I think people would lead an impoverished life if their life plans aimed only at their own well-being, and yet the main principle requires only that improvements to the structure of law and morality equitably promote well-being. Why doesn't the main principle include any of the other values that would make their lives worthwhile?

On this question, I am persuaded by Scanlon's example of the person who chooses to fast to save money to build a shrine to his god (1975). The main principle leaves each person free to pursue whatever values he thinks are important in a worthwhile life, but it does not require that the cooperative structure of society equitably promote those values. No matter how important it is to someone to build a shrine to his god, there is no social responsibility to promote that goal. If the social structure provides an equitable

distribution of opportunities for well-being, people are free to use those opportunities to promote whatever values are important to them. In that way, they are able to lead worthwhile lives.

Indeed, even to say that the main principle endorses the equitable distribution of well-being gives it too much credit. Recall that in chapter 4, I distinguished between the broad and narrow conceptions of well-being. On the broad conception, respecting the rights of others would be part of a good life. However, to include respect for the rights of others in my account of well-being would make it trivial to say that promoting well-being requires respect for the rights of others. This is the circularity problem raised by R. Dworkin (2000). In chapter 4, I avoided this circularity problem by assuming a narrow conception of well-being. The narrow conception of well-being enables the main principle to provide a noncircular explanation of human rights, but at the cost that there is no logical guarantee that equitably promoting well-being in the narrow sense would even be a good way of promoting well-being in the broad sense. That it is a good way of doing so is due to the fact that, given a suitable background structure, normal adults are able to use opportunities for well-being in the narrow sense to promote all their values, including well-being in the broad sense.

Are there any values other than well-being? In previous chapters, I have shown how to explain a number of other values—for example, the value of choice—indirectly in terms of well-being. Pettit (1997) proposes that governments should take as their goal the minimization of *domination*—roughly, a relation in which one person can arbitrarily exert his will over another person. Now it is clear that domination can adversely affect well-being. So Pettit has not picked a value that is orthogonal to well-being. Pettit's main argument for distinguishing between the badness of domination and the badness of reductions in well-being is that that you can imagine being the slave of a kindly master (1997, 22). Even though the master never does anything that adversely affects your well-being, you are still his slave and that would be bad.

I think there is a puzzle with this argument. Suppose, for example, you lived your entire life never knowing that you had a master. Your master had the legal right to force you to do anything he wished, but he wished only for you to live your life as you saw fit, so he never exercised his power and never informed you of it. Would your life be worse because he had the power? Well, maybe a little, but not much. Certainly there are lots of things that are much worse. In any case, explaining the badness of slavery, even with a kindly master, is not a problem for the indirect consequentialist, who can easily explain the badness of the *practice* of slavery even if some slaveholders are kindly.

No welfare consequentialist in my sense of the term would ever recommend minimizing domination if it were costly to do so, because reductions in domination would have to be balanced against other potential improvements in life prospects. Would we cancel the federal budget for cancer research so

that the money could be used to reduce domination? Doesn't it depend on how bad the domination is? It seems to me that the government should focus its attention on the kinds of domination that adversely affect life prospects.

Finally, I don't think Pettit is right that domination is always bad. Every night lots of people go to sleep comforted by the belief that they are dominated by a benevolent God. They may be mistaken about the facts, but I don't see that they are mistaken in thinking that some kinds of domination would not be bad.

Just as it is important to emphasize that I don't think that a life devoted entirely to one's own happiness is a very worthwhile kind of life to live, I don't believe that well-being is the only value. Thus, it is something of a surprise that it is the value by reference to which improvements in morality and the law are evaluated.

Humans, Angels, and Demigods: The Second Contingency Objection

Perhaps the most compelling objection to any consequentialist defense of human rights is that it makes human rights contingent on their consequences, in my case on their equitably promoting life prospects when evaluated as a practice. But human rights seem to be more fundamental than that. They seem to express unconditional moral demands that each of us is entitled to make simply in virtue of being human. This is a different objection from the trade-offs objection. It raises an issue of how intimately the concept of human and human right are connected.

One part of this objection was addressed in the first volume (Talbott 2005, 185–186). There I imagined a society of beings like us, except that autonomy made them miserable. I argued that it would be cruel to require that they all develop their autonomy rights if it were possible to arrange society so that only a few would have to develop their autonomy rights (and be miserable) so that they could take care of the others, who would lead happy lives. In such a society, guaranteeing universal human rights would make everyone miserable, and so, it would not be endorsed by the main principle.

There is another way of thinking about this contingency objection. In chapter 5, I explained the paradox of direct consequentialism and used it to explain why the main principle does not endorse our using the main principle as a ground-level moral principle. As I explained it, ground-level human rights norms are our indirect way of satisfying the main principle. So it seems that human rights norms are appropriate for us because of our cognitive limitations. For cognitively superior beings, they would be irrelevant.

For example, what about angels, whom I imagine to be just like humans, except that they always make a good faith effort to determine what is right and then they do it? Angels would have the same limitations as humans in

being able to foresee the effects of acts and practices in equitably promoting life prospects, so I don't see how they could avoid the paradox of direct consequentialism either. They would still need a legislature to enact laws and to improve them, and they would still need some kind of judicial system to resolve conflicting interpretations of the law and to develop improved interpretations of it. I think they would still need legally established rights, not because they would need the law to protect those rights, but just to coordinate expectations and behavior. They would not need police or prisons or other institutions for law enforcement. For them, Rawls's ideal of strict compliance would be a reality.

What kind of beings would be able to dispense with rights? It would have to be a being with the conscientiousness of angels and with godlike powers of foreknowledge. Let's call them *demigods*. Demigods could foresee the effects of alternative practices on life prospects and could also envision new kinds of practices that we humans will never imagine. Demigods would be able to apply the main principle to alternative practices to determine which would do the best job of equitably promoting life prospects. The only laws they would need would be laws to break ties among alternative practices. For example, if the practice of driving on the right and driving on the left had equivalent effects on life prospects, then they would need to establish a convention so that everyone would know what to expect everyone else to do. They would almost surely not need the concept of a right to know how to treat other demigods in a way that would do the best job of equitably promoting life prospects. And so, for demigods, rights would almost surely be superfluous.

Does the fact that rights would almost surely be superfluous for demigods somehow diminish their importance for human beings? Or does it rather somehow diminish the significance of human beings? I don't see why either question should be answered in the affirmative. If understanding ourselves as the bearers of human rights is our best conceptual framework for equitably promoting the life prospects of human beings everywhere, that would seem to be a strong endorsement of human rights.

But it *would* be some kind of logical limit on their universality. Human rights would apply to all human beings and even to all angels, but they would not really be appropriate for demigods. Perhaps this comes as a disappointment.

Maybe there is another worry lurking. Suppose the demigods were to look down on human beings and human practices and ask themselves this: What kind of practice for human beings would do the best job of equitably promoting their life prospects? Would their answer be that human beings should employ a human rights framework? Or would they see a superior conceptual framework for moral thinking that would do a better job of equitably promoting human life prospects?

Suppose they did see a superior alternative. Suppose also that they could hypnotize human beings to adopt this new moral framework. When humans

woke up from the hypnosis, they would have all the capacities they currently have; only their consciences would be different. We, of course, can't imagine what their consciences would be like. However, by hypothesis, human beings would no longer think of each other as the bearers of human rights, but as having some other moral status.

The thought experiment does not work if you imagine that posthypnosis, humans are like zombies following the will of the hypnotists. We know enough about human well-being to know that that would not be a good life for human beings. What you have to imagine is that the hypnosis was really just a quick way of producing the effects that, in humans, are typically produced by moral training. We cannot suppose that the change would make human beings perfectly moral beings like angels. Humans would still have their moral shortcomings. It is just that, shortcomings and all, the new moral practices would be favored by the main principle because of their effects on the equitable promotion of life prospects.

Would it be permissible for the demigods to perform mass hypnosis on human beings to change us over to the new moral framework? Wouldn't doing so be a violation of our human rights? Technically, it would be an infringement of our human rights, but given that they had the knowledge to be able to rule out the possibility of bad side effects, the main principle would endorse a moral system for *them* that permitted such infringements.

The idea of a radical restructuring of our moral thought is not a new idea. Both the French Revolution and the Russian Revolution announced the creation of a new kind of man with a new kind of morality. When human beings set themselves up as demigods in this way, the results are always disastrous. We need a right to freedom of conscience as protection against this kind of *hubris* by other human beings.

But would we need the right to protect us from true demigods? Suppose tomorrow you woke up and discovered that you and the rest of humanity had a completely new way of thinking about right and wrong that did not involve rights. And suppose that this new way of thinking about right and wrong greatly transformed human life. War, starvation, malnutrition, and torture were almost completely eliminated. When the demigods took over our TV stations to announce what they had done, would you feel that they had violated your rights? Would you think you had a claim against them for violating your rights? What kind of compensation would they owe us for eliminating such evils from our lives? Would you think that they had a moral obligation to undo the hypnosis and change us all back to our former selves—send us back to a world of wars, starvation, malnutrition, and torture?

This is, of course, a total fantasy. When we come back to the real world, it is important to appreciate the role of human rights in equitably promoting life prospects. The human rights framework really is our best hope of eliminating war, starvation, malnutrition, and torture. It won't happen overnight. But it is happening.

Could Utilitarianism Be True After All?

In this book, I have defended an indirect consequentialist principle of moral improvement, the main principle, a principle of equitable division of the benefits and burdens of cooperative practices. I have emphasized that this principle is not utilitarian, because it pays attention to the distribution of well-being. But now a worry arises. Perhaps the main principle is not the most fundamental principle of moral improvement. Recall that I criticized contractarian theories on the grounds that they are not the most fundamental explanatory level. There is a lower level at which the main principle explains why some agreements are reasonable and others are not.

But now the question arises: Is the level of explanation invoking the main principle the most fundamental level of explanation? Perhaps there is a deeper level of explanation, at which the explanation employs the utilitarian principle. How could this be? Here is one possibility: Perhaps the main principle is a limited principle that applies only to beings who are so selfish that they cannot be effectively motivated to directly aim at the goal of maximizing overall well-being (as required by the utilitarian principle). It is easy to imagine that it is part of our evolutionary legacy that we give our own well-being and the well-being of our children more weight than we should. Members of earlier generations whose moral judgments were more utilitarian would have been less likely to reproduce, so there are few, if any, of those people around today. Given the kind of people that evolution has produced, the only stable human societies will be those in which individuals can think of themselves as part of a cooperative scheme in which the benefits and burdens of cooperation are shared fairly.

If these factual claims were true, the utilitarian principle would endorse the main principle as a special case. The moral appropriateness of the main principle as a meta-level principle of ethics would be explained at an even higher level of abstraction by the fact that, for human societies, satisfying the main principle is the best way of maximizing overall well-being.

So, it seems that utilitarianism might be true after all. How might we evaluate this kind of very indirect utilitarianism? It is not easy to see how to do so. It will not do to simply echo Rawls's dictum that "utilitarianism does not take seriously the distinction between persons" (1971, 27), because this new version of utilitarianism proposes to explain why we take the distinction between persons so seriously and to explain why we *should* do so. However, I think Rawls's idea can still be defended.

To begin with, we should be suspicious when a utilitarian tells us that it just so happens that the facts in our world are such that utilitarianism agrees with our considered moral judgments. In chapter 4, I referred to this as an *actual-world narrowing* of the theory. We should be suspicious of actual-world narrowings of utilitarianism or of anything else.

To see if this defense of utilitarianism is an actual-world narrowing, we would need to consider other worlds, different from the actual world, to help

us decide whether, from a moral point of view, it is only the (total or average) amount of well-being that is morally significant, or whether its distribution is also morally relevant. Let us consider a society of unselfish beings who, unlike us, are willing to sacrifice themselves for the good of the whole. Let us suppose that there are two alternative systems of social practices that are potentially the best for them. In one system, well-being is distributed relatively equitably. Call it system E. In the other, well-being is distributed quite inequitably. Call it system U. For example, in U, a subclass of members of the society are trained from birth to be miners and they spend their lives working in the mines. The rest of society engages in less arduous pursuits that leave them plenty of leisure to pursue the arts and philosophy.

In contrast, in E everyone spends 5 years working in the mines so that everyone may spend the rest of their lives in less arduous activities that leave them the leisure time to pursue the arts and philosophy. As it happens, U is more efficient, so that average and total life prospects are higher in U. Thus, according to utilitarianism, U is morally superior to E. Just to make the case for U as strong as possible, let us suppose that the miners in U don't complain about their lot. Because they value well-being per se, though they think it is unfortunate that anyone has to be a lifelong miner, they willingly perform their role for the good of the whole.

From a moral point of view, is U better than E? How are we to decide? I see no alternative to thinking carefully about the case and arriving at a considered moral judgment. When I do think about it, I can't find any moral significance in the fact that total and average well-being are higher in U than in E. It seems to me that E is morally superior to U, because in U the miners' life prospects would be much lower than anyone's life prospects in E and everyone's life prospects in E would be pretty good. So I do not think that the utilitarian principle explains the main principle, even if we assume that the empirical facts are such that the main principle would be endorsed by the utilitarian principle as a principle of moral improvement for this world.

I follow Rawls in thinking that most moral practices are, at the most fundamental level, practices aimed at securing the benefits of social cooperation and equitably distributing their benefits and burdens. It might seem that in a world comprised entirely of impartial utilitarians, whose only goal was to maximize overall well-being, there would be no role for this sort of morality. This is a mistake. Even if they all agreed on the goal of maximizing well-being, in any world in which individual agents could disagree about the probabilities of relevant outcomes, there would be a potential for collective action problems and thus a role for moral practices that generate solutions to collective action problems that equitably distribute the benefits and burdens of cooperation.

Consider an example. Suppose there were only two countries, one with a capitalist economy (CAPITAL) and one with a socialist economy (COMMON). The leaders of CAPITAL believe that establishing capitalism worldwide is the way to maximize overall well-being. The leaders of COMMON

believe that establishing socialism worldwide is the way to maximize overall well-being. It is easy to see that their conflicting beliefs about how to maximize well-being could lead to an arms race, a classic collective action problem. In such a case, both countries might well be better off if they could rely on each other to make and abide by an arms control agreement than they would be if they engaged in an arms race. The practice of making and abiding by agreements could promote well-being in both countries by facilitating possible mutually beneficial cooperation. If so, the development of such a practice would be favored by the main principle. Thus, perhaps surprisingly, the main principle, and the conception of morality that it underwrites, could have an important role to play even in a world of pure utilitarians.[1]

Conclusion

Evolutionary biologists can explain why evolution would favor social beings who could recognize one another and engage in reciprocal cooperation. One species of such beings developed language and the ability to share reasons. At some point their reasoning became sensitive to objective reasons—not so sensitive that they were infallible, but sensitive enough so that over time, their reasoning tended to improve itself. This process took place in every area of reasoning: reasoning about what to believe was sensitive to truths of epistemic rationality; reasoning about what to do in nonmoral situations was sensitive to truths of individual rationality; moral reasoning was sensitive to truths of reasonableness. Because the third kind of reasoning employed empathy, most philosophers did not even recognize it as a form of reasoning. They kept trying to make moral reasoning fit the model of the second kind of reasoning, to be a kind of individual rationality. It might never have been recognized that moral reasoning involved sensitivity to objective truths, were it not for the fact that there emerged from the process of its development the most elegant principles, principles of maximizing overall happiness or of impartial agreement, principles that were clearly universal, not parochial.

And what was the process that led to the emergence of these principles? It was the same kind of process that had led to improvements in each kind of reasoning over time: a more or less free give-and-take of opinion, giving and responding to reasons. It turned out that this process had the potential to make the group more sensitive to these reasons than its individual members. For that potential to be realized, it was necessary to have as many different voices contributing to the free give-and-take of opinion as possible.

Over time there developed a conception of robust and inalienable basic human rights, of two kinds: first, the autonomy rights that, if guaranteed to everyone, would enable them to contribute to the process of the free give-and-take of opinion; second, political rights, that guaranteed that social decisions would be sensitive to the feedback generated by the free give-and-take of opinion.

The process did not stop there. Over time, as the process of free give-and-take of opinion led to increased moral sensitivity, more rights were added to the list of human rights. In this book, I have done my best to project where the process is going and to answer the question with which I began my first volume.

Which Rights Should Be Universal?

What are the human rights that should be universal? I begin with the basic human rights, but now expanded to include the procedural rights that are included in the right to physical security (chapter 6) and the various rights against paternalism (chapter 13):

1. A right to physical security, including procedural rights such as due process of law
2. A right to physical subsistence (understood as a right to an opportunity to earn a subsistence for those who are able to do so and a welfare right for those who are not)
3. Children's rights to what is necessary for normal physical, cognitive, emotional, and behavioral development, including the development of empathic understanding
4. A right to an education, including a moral education aimed at further development and use of empathic understanding
5. A right to freedom of the press
6. A right to freedom of thought and expression
7. A right to freedom of association
8. Liberty rights to a sphere of personal autonomy free from legal paternalism, including the following:
   a. A right to religious freedom
   b. A right to sexual freedom
   c. A right to reproductive freedom
   d. A right to refuse medical treatment, including a right to refuse extraordinary care and to be removed from life support
   e. A right to marry that includes same-sex marriage
   f. A right to suicide and assisted suicide in certain end-of-life situations
9. Political rights, including constitutional protections of the human rights, a democratic procedure for adopting legislation, and an independent judiciary to interpret and apply the constitution.

The nonbasic rights are these:

10. Economic rights, including property and contract rights in a regulated market economy, rights to gainful employment, to unemployment compensation, to minimum wage (including prohibitions on slavery and indentured servitude), to occupational health and safety, to collective bargaining, and to bankruptcy protection
11. Negative opportunity rights—that is, rights to protection from systematic discrimination, including discrimination on the basis of race, color, sex, national or ethnic origin, religion, age, and disability, sexual orientation (and perhaps height, weight, and appearance)

12. Positive opportunity rights—that is, in addition to the development of judgment rights 1–4 above, development of one's capabilities to assure a reasonable choice of careers; this includes persons with disabilities, if they have the necessary capacities
13. Rights to social insurance, including rights to disability insurance, health insurance, retirement insurance, maintenance insurance, including food, housing, and welfare
14. Privacy rights, including the following:
    a. A right to a private space, protected from physical intrusion and certain other kinds of access
    b. A right to informational privacy, that certain kinds of content be protected from being revealed

Comparison with the U.N. Universal Declaration of Human Rights

If we compare my list of human rights with the rights in the U.N. Universal Declaration, we find that almost all of the U.N. rights are on my list. The only ones missing are those that are too specific to be human rights, such as the right to holidays with pay. However, my list is more expansive than the U.N. Universal Declaration. Some of the additions are rights that have been included in later human rights documents, including rights to reproductive freedom and rights against discrimination on the basis of disability or sexual orientation.

However, some of the items are not found on any current human rights document, though they are included in some constitutions, including these: a right to sexual freedom, a right to same-sex marriage, and rights to make end-of-life decisions, including the right to refuse medical care and to assisted suicide.

This is an expansive list. There is no country in the world that guarantees every right on the list, and there is no prospect of an overlapping consensus on many of them in the foreseeable future. Why include rights that are only aspirational at this time? For as long as there has been the concept of human rights, human rights have been aspirational. There is nothing on my list that is more aspirational than a right against slavery was in the eighteenth century. And yet, the claim that there was a right not to be enslaved played an important role in developing the consensus that eventually eliminated it. I refer to this as the *prescriptive* role of human rights. It is an important role that human rights discourse has played historically and that it continues to play.

Still, the list is a long one. Is it too long? When the U.N. Universal Declaration was adopted, it seemed a hodgepodge. There was no obvious rationale for the items on the list, which was politically advantageous because it would

not have been possible to obtain consensus on a rationale. But the rights on my list are not like that. We can understand their rationale. If, as I have argued, sensitivity to the equitable promotion of life prospects has played a role in the process of moral and legal improvement for thousands of years, then we can expect that there will come a day when all the rights on my list will be generally recognized. When that day comes, my list will not seem too long, but too short.

The End of Reasonable Disagreement?

What good is a list of human rights if there is reasonable disagreement over the items on the list? Well, there is reasonable disagreement over just about everything in philosophy. From the point of view of the Proof Paradigm, that is a big problem, because there could not be reasonable disagreement about something self-evident or provable from self-evident premises. Philosophy and political theory are not like that. They are full of reasonable disagreement. We don't make progress by ending reasonable disagreement. It is the engine of progress.

Suppose I were to publish this book and when the reviews appeared, the reviewers found nothing to disagree with. Then suppose that the book was adopted in courses on human rights throughout the world, and when it was discussed in those courses, everyone agreed with everything in it. This sounds like a philosopher's dream. Mill was the first philosopher to recognize that it would be a nightmare.

What could possibly explain the fact that no one disagreed with anything in the book? Not that everything in the book was true. I am as sure as I am of anything that that is not true. I could conclude only that, due to some sort of influence I did not understand, people everywhere had lost their capacity for independent judgment. This would be terrible news, because it would signal the end of the historical-social process of moral discovery that I had intended to be contributing to. The process of free give-and-take of opinion that had been driving moral progress for thousands of years would have ground to a halt, and moral progress with it. So please, right now, think of something you disagree with in this book. Thought of something? What a relief! There was nothing to worry about. The process continues.

Notes

Chapter 1

1. Here I intend to be acknowledging only what seems to me obvious: that some human adults are too impaired to be accorded all the rights on my list of human rights. This does not mean that they would have no rights, only that their rights would be different. How are we to draw the line between normal and nonnormal for the purposes of assigning human rights? Ultimately, as I explain in chapter 13, I draw the line in terms of my consequentialist conception of autonomy. I note here only that, given the history of abuses in categorizing groups as less than fully human, it is extremely important to insist that anyone who categorizes another person or group as not normal, and thus not due the full complement of human rights, assumes a substantial burden of proof.

2. It should be noted that, based on Mill's account of the writing of *On Liberty* in his autobiography (Mill [1873]), *On Liberty* was really a coauthored work. Mill's coauthor was his wife, Harriet Taylor Mill.

3. Mill seems to be advocating absolute liberty rights when he says, "No society in which these liberties are not, on the whole, respected, is free, whatever may be its form of government; and none is completely free in which they do not exist absolute and unqualified" ([1859], 19). However, in other places, Mill acknowledges that "rules of conduct cannot be so framed as to require no exceptions . . ." ([1863], 299) and goes on to list reasons why it is hard to avoid some exceptions. In any case, he himself allows exceptions to his autonomy rights—for example, he allows a limit on freedom of expression in the case in which the opinion that corn dealers are starvers of the poor is expressed to an angry mob outside the home of a corn dealer ([1859], 64). An even more important exception to his liberty rights is his refusal to permit slavery contracts ([1859], 115).

4. Although there are nonconsequentialists, such as Kant, who admit of no exceptions to rights (e.g., Kant [1797]), they are very much the exception among nonconsequentialists. Even the early Nozick allowed for exceptions to rights to avoid "catastrophic moral horror"(1974, 30n; see also 1981, 495).

5. In the literature, there is nothing like unanimity on what makes a normative theory *consequentialist*. *Consequentialism* is sometimes defined more narrowly than I have defined it, to require that evaluations be based on the nonmoral value of *states of affairs*. This is too narrow a notion for my purposes, because I want to allow for the well-being of a life to be a value of the life as a whole, without assuming that it can be decomposed into a sum of the

values of individual time slices. Also, it is often assumed that a consequentialist view must be a maximizing view. However, I allow for distributive considerations to enter into the evaluation. It may seem that the maximizing constraint is trivial, because given any complete ranking, it is possible to define a function that the ranking maximizes. However, I wish to leave open whether there is a complete consequentialist ranking based on well-being or whether the ranking is only partial, as it would be if some alternatives were incommensurable. See Raz (1986, chap. 13). I say more about my use of the term *consequentialist* in chapters 3, 4, and 14.

6. *Metaphysical Rawls* is my reconstruction. I have no way of knowing if he ever existed, but I suspect that he represents a side of the early Rawls that was present in 1971, but disavowed later. For a different and equally cogent reading of *Theory of Justice*, see Freeman (2007).

7. After this introduction, Rawls tempered his statements by acknowledging, "No doubt they are expressed too strongly" (1971, 4). But by then he had already achieved his rhetorical purpose of separating himself from Mill and the other utilitarians. I should note that the rhetorical force of Rawls's introduction was so great that it obscured the fact that what Rawls says about justice does not seem to be true of the ordinary conception of justice. For reasons given by G. A. Cohen (2008, chap. 7) and that I discuss in chapter 3, I don't believe that Rawls's theory of *distributive justice* is a theory of justice at all. Nonetheless, in this book, for the sake of being understood, I will often refer to Rawls's theory as a theory of *distributive justice*. That distributive justice is not really justice will not come as a surprise to those who are familiar with the way names work in language. After all, American Indians are not from India either.

8. Because the early Rawls's general conception of justice is explained in terms of what would be chosen in an original position, it might be argued that his account is at the most fundamental level an account in terms of hypothetical choice, and thus is nonconsequentialist. As I explain in note 12 below, I think this is a mistake.

9. Rawls later made some revisions to his statement of the principles. I quote the revised principles here:

a. Each person has an equal right to a fully adequate scheme of equal basic liberties which is compatible with a similar scheme of liberties for all.
b. Social and economic inequalities are to satisfy two conditions. First, they must be attached to offices and positions open to all under conditions of fair equality of opportunity; and second, they must be to the greatest benefit of the least advantaged members of society(1993, 291).

10. Mill himself advocated a representative democracy on utilitarian grounds, but favored plural voting rather than one person, one vote ([1861], chapter 8). It should also be noted that the autonomy rights on Mill's list are characterized much more broadly than the corresponding rights on Rawls's. The most significant difference is that Mill includes a broad right of normally functioning adults to be free from paternalism, a right that is largely missing from Rawls's list. I discuss this right in chapters 12–13.

11. "All social primary goods—liberty and opportunity, income and wealth, and the bases of self-respect—are to be distributed equally unless an unequal distribution of any or all of these goods is to the advantage of the least favored"(1971, 303). Rawls himself thinks of primary goods as part of a "thin" (1971, 396) conception of the good. The theory must be "thin" enough not to include any moral components. Thus, Rawls's principle qualifies as consequentialist in my sense. Because Rawls does not present primary goods as proxies for well-being, his theory is not a welfare consequentialist theory. However, all that is necessary to make the early Rawls's theory welfarist is to interpret Rawls's "thin" theory of the good as a "thin" theory of well-being, as Rawls himself seems to do at some places (e.g., when he says that "the index of well-being" is specified in terms of primary goods (1971, 396). This we must be able to do if Rawls's "thin" theory is adequate, because to be adequate, the primary goods must be things that would be useful in pursuing any rational life plan. Surely, many, if not all, rational life plans include provisions for one's own well-being.

12. Because the early Rawls's general conception of justice is explained in terms of what would be chosen in an original position, it might be argued that his account is at the most fundamental level an account in terms of hypothetical choice, and thus is nonconsequentialist. I believe that this is a mistake. Because Rawls's reflective equilibrium method leads him to adjust the original position to fit our considered moral judgments (1971, 21, 141; 1993, 25–28, 275), if those considered moral judgments have a consequentialist explanation, then so does the original position construction. Therefore, if Rawls's original position thought experiment yields consequentialist basic principles of justice, then I regard the entire theory as consequentialist. Note that nothing crucial hinges on this interpretive issue. Even if it is a mistake to think that the early Rawls's general theory of justice was consequentialist, his most basic principle of justice, the maximin expectation principle, is.

13. An exception is Joseph Raz's (1986) defense of the importance of autonomy as part of a good life, at least in societies like ours. Raz's account is very different from mine, but it is not incompatible with it. Raz's insights into the importance of autonomy for the good life for an individual and the importance of social forms for realizing the autonomous life complement my discussion of autonomy as a social achievement and of the contribution of autonomy rights to the social project of equitably promoting everyone's well-being. However, Raz's account is narrower than mine, because his account attempts to explain only the direct value of autonomy to the autonomous individual in an autonomy-enhancing culture (1986, 390–391), whereas my account attempts to explain why, when combined with democratic rights, autonomy rights indirectly promote well-being in any culture. My view is more closely related to the capabilities approaches of Martha Nussbaum (2000) and Amartya Sen (1999, 2000, and 2009). What is missing from their views is an explicit connection between rights to capabilities and well-being. I discuss Nussbaum and Sen more fully in chapter 4. I should also mention that Pettit (1997) counts his view as consequentialist because it is a maximizing view. He would maximize nondominance. I am not sure how to classify his view. It would count as consequentialist in my sense if he can explain

dominance in nonmoral terms. I discuss Pettit's view in chapter 14. Finally, this is only a partial list. If I were to try to list all the nonconsequentialist approaches to justice and rights in the literature, there would be over 100 names on the list. The number of consequentialist accounts would be only a small fraction of that number.

14. Bottom-up reasoning can also lead to giving up a norm or principle, as when, for example, I give up the norm that killing a human being is always wrong when I decide that killing in self-defense is sometimes justified. Because I allow ground-level moral reasoning to be either top-down or bottom-up, I follow Rawls (1971, 48–50) in characterizing it as a kind of *equilibrium* reasoning. For more on these distinctions see Talbott (2005, chap. 2).

15. LeBar (2009) makes a similar criticism of Darwall, though not in defense of consequentialism, but in defense of virtue ethics. LeBar distinguishes between the second-personal *content* of our moral reasons (my ground-level reasons) and the justification for our dispositions to respond to reasons with second-personal content (my meta-level explanatory reasons). I would simply add to LeBar's argument that there is a higher meta-level social practice consequentialist explanation of why the traits and attitudes (i.e., virtues) that he would justify by reference to the good life make a life good. A good life for human beings consists of moral and nonmoral elements. I believe that the moral elements are primarily the traits and attitudes that are part of social practices that equitably promote well-being (in the narrow sense), or at least do a better job of equitably promoting well-being than any of the relevant alternatives

16. I take the term *practice* from Rawls, but not his definition, because as Rawls defines it, a practice is "a form of activity specified by a system of rules" (1955, 3 n. 1). As I use the term, it applies to any structured form of activity, whether specified by rules or transmitted in some other way (e.g., by training). A *social* practice is a practice that requires multiple participants—for example, the family.

17. Note that my primary/secondary distinction is not the same as Hart's (1961) well-known distinction in the law.

18. The most important difference between this volume and the first is that in the first volume I suggested that our justification for accepting human rights norms depended on our being *epistemically justified* in believing that human rights norms satisfied what I am now calling the *main principle* (Talbott 2005, 116). I now believe that that was a mistake. The main principle is an objective principle of moral improvement, not a subjective principle that we apply in ground-level moral reasoning. As I explain in chapter 5, I now believe that good moral reasoning does not typically involve explicit application of the main principle, but rather implicit sensitivity to it.

19. The two main categories of the Golden Rule are "Do unto others as you would have them do unto you" and "Love others as yourself," though each category has different variations—for example, others may be referred to as "your neighbor" (Judaism and Christianity) or "your brother" (Islam). The various versions of the Golden Rule are quite useful ground-level principles, but, like all useful ground-level principles, they have exceptions, as I discuss in note 23 below.

20. Perhaps the first occurrence of the Golden Rule as a centerpiece of ground-level moral thought is in the Analects of Confucius (551–479 BCE), but it seems to have been independently discovered a large number of times. At the 1993 Parliament of the World's Religions, 143 representatives of all the world's major religions signed a statement endorsing the Golden Rule as part of a Declaration of a Global Ethic (1993).

21. Strictly speaking, the ground-level moral practice applies to all of judicial and legislative determinations, because the ability of judges or legislators to distinguish between cases that raise issues of justice and cases that do not is itself part of the ground-level moral practice.

22. Indeed, as I mention shortly and discuss more fully in chapter 5, there is good reason to think that, in most cases of interest, no finite specification of the facts of an example *could* rule out all exceptions.

23. Of course, there are well-known exceptions to the Golden Rule itself—for example, the masochist who would insist that it is morally permissible for him to cause pain to others because he would like them to cause him pain; or the extremely competitive racer who insists that you (who are leading the race) have a duty not to cross the finish line ahead of him, because if the roles were reversed, you would not want him to cross the line ahead of you. There are an endless variety of exceptions, which, as I explain in chapter 5, is important evidence for the truth of the main principle.

24. For ease of exposition, I will generally refer only to the positive project of explaining the parts of ground-level moral thought that are true or appropriate and I will take for granted the qualification "when they are true or appropriate." However, the negative project of explaining those parts of ground-level moral thought that are false or inappropriate (when they are false or appropriate) is understood to be included.

25. In the previous volume, I explained why not all moral views are equally valid (Talbott 2005, chap. 3).

26. In Talbott (2005), I explained that I am a moral realist of a particularly strong kind: I believe that the fundamental principles of morality (i.e., the moral meta-principles) are true in all possible worlds. I also believe that we can make reliable (but not infallible) moral judgments about particular cases. In this book, I further develop my moral metaphysics and epistemology, but I do not attempt any sustained response to moral antirealism or moral skepticism.

27. A collective action problem is a situation in which, even if everyone in the relevant group acts rationally, the outcome may be worse for everyone in the group in their own estimation than it would have been if they had all chosen differently (M. Taylor 1987, 19).

28. To classify these examples as an N-Person PD is an oversimplification, because members of social groups do not interact with one another only one time; they interact many times over the course of their lives. However, the simplification permits me to easily characterize cooperating, defecting, and free riding without the complexity that a more rigorous analysis would require.

29. In the previous volume, I gave an example of cockroach people to illustrate the difference between moral realism and moral antirealism (Talbott 2005, 169–170).

30. This is not the only way that the main principle differs from utilitarianism. I discuss the features of utilitarianism that doomed it as an explanatory meta-theory in chapter 3.

31. Hume famously claimed that justice would not be a virtue in a world of abundance ([1777], section 3, part 1). This is a mistake. Competitive goods (e.g., being the fastest sprinter in the world) are logically scarce. Even in a world of plenty, there would still be lots of CAPs. Indeed, as I explain in chapter 14, even if everyone in the world wanted to do nothing but maximize overall utility, there would still be a potential for CAPs based on disagreements about how best to do it. So probably the only way for rational beings to avoid CAPs is to have no goals, which may explain why so many Eastern religions characterize enlightenment as freedom from all attachments. However, freedom from all attachments is a high price to pay to avoid CAPs.

32. In the previous volume, I explained the minimum standard of moral legitimacy in terms of what is necessary for a government to have a right against intervention, whether coercive or noncoercive. I now think it is better to explain moral legitimacy in terms of what is necessary for a government to be recognized as having the moral authority to act for and bind its citizens (cf. Reidy 2005). I should also note that by saying that the basic human rights are rights of normal human beings, I do not mean to imply that nonhuman beings or human beings with disabling cognitive, emotional, or behavioral impairments should have no rights, only that their rights are different. An important practical question is the question of where, for human rights determinations, to draw the line between normal and nonnormal. I take up this question in chapter 13.

33. Here I use *autonomy* in a nonmetaphysical, consequentialist sense that I explain in chapter 12.

Chapter 2

1. Would a pacifist disagree? It depends on the kind of pacifist. As I understand it, Moses, Fred, and Bob have discovered a moral permission, roughly, that it is permissible to use force in self- or other-defense. I do not believe that using force in self- or other-defense is morally required. One way to be a pacifist is to agree that self- and other-defense is morally permitted, but to choose not to engage in it. This kind of pacifist can agree with me on the example. A pacifist who believes that self- or other-defense is not even morally permissible would disagree with me on the example. Even pacifists of this kind should recognize an important moral difference between Adolph's coercion and Winston's.

2. Hayek (1960) gives up on distinguishing justifiable from unjustifiable coercion, and adopts the principle that coercion should be minimized. This seems to me to be a mistake, which is why I track the account of Nozick (1974) in my example.

3. In chapter 6, I specify the contents of the list of basic harms more precisely.

4. In my account of libertarian rights, I have followed Nozick (1974) in supposing that there is a natural right against an imposition of a risk or threat

of harm, and thus a right against coercion. Somewhat surprisingly, Thomson disagrees (1990, 244). It would take me too far afield to consider the issues between them here. Suffice it to say that Nozick's example of someone who plays Russian Roulette on you against your will (1974, 74, 79) strikes most people as a serious rights violation, as evidenced by our willingness to permit coercion to stop them from doing it in the present, to deter them from doing it in the future, and to punish them for doing it in the past.

5. Locke's [1690] is the traditional account of natural rights against basic harms. It included an enforcement provision covering all the elements listed here: prevention and deterrence of potential rights violations and compensation for and punishment of actual rights violations. Note, however, that enforcement rights carry their own enforcement provisions, which in turn generate an unbounded hierarchy of enforcement rights. For example, if I violate your natural rights and cause you a basic harm, you have a right to compensation from me for the rights violation. If I cause you basic harm when you attempt to collect that compensation, you have a right to further compensation for the additional harm you suffered in trying to obtain compensation for the initial rights violation. Obviously, there is no theoretical limit to the rights to compensation that can be generated in this way.

6. This is my best recollection of the example Nozick used when he explained the exception in the course I took on his book manuscript in the fall of 1972.

7. The example is a variation on Nozick's (1974, 181). It is not so far from reality as one would hope. When Alexis St. Martin appeared at the door of William of Beaumont with a life-threatening gunshot wound, William Beaumont offered to perform surgery only if St. Martin would agree to an unusual kind of indentured servitude: St. Martin had to allow Beaumont to maintain physical access to his viscera in order to carry out studies of gastric physiology (Veatch 1987, 208).

8. To see this, consider the deeper principle that Nozick himself appeals to as a justification for libertarian rights: Kant's categorical imperative to never treat others as means only, but also as an end (1974, 30–31). No reasonable interpretation of that principle could justify Marie's using her bargaining position to enslave everyone else on earth.

9. By *substantive* norms and principles, I mean norms and principles that would be useful for helping us to make moral decisions in real-world cases. Thus, I exclude such principles as "Do the right thing," which is of no guidance at all, and also such principles as "It is always wrong to torture children for the fun of it," which no one with any moral sensitivity would ever need to make use of.

10. Of course, Rawls limits the deliberation in the original position to the principles of justice for the basic structure of society (1971, 7). In this book, I use a variation on Rawls's construction without this limitation. I discuss the construction and my use of it more fully in chapter 4.

11. Note that it will not help to say that what makes a process of real-world discourse (RWD) a good approximation of the ideal process of rational discourse (IRD) is that if the participants in the IRD were to consider the question, they would agree that RWRD is a good approximation of IRD. I discuss Habermas's theory more fully in chapters 7 and 10.

12. *Williams v. Walker-Thomas Furniture Co.*, 350 F.2d 445 (D.C. Cir. 1965).

13. Model Penal Code, Section 3.02 (American Law Institute 1985). Note that this statement is clearly inadequate. It immediately gives rise to lots of exceptions, even if it is qualified to require, as it must, that the harm avoided be much greater than the harm of breaking the law.

Chapter 3

1. Note that what I am calling a *coordination problem* here is a relative of the problem addressed in Regan (1980). A full solution to this coordination problem requires the use of an equilibrium decision rule, as I explain in Talbott (1998). I set aside this potential complication here.

2. The example captures some of the features of *Regina v. Dudley and Stevens*, 14 QBD 273 DC (1884), in which four shipwrecked crew members who were dying of starvation and thirst in a lifeboat discussed throwing dice to determine who would be eaten by the others. No dice were used because Parker, the cabin boy, went into a coma and so he was eaten by the other three. In this case, although the law allowed no necessity defense to murder, popular opinion and the jury were so much on the side of the defendants that to secure a conviction, the judge had to write the jury's verdict himself.

3. I here side with those who think there is a prima facie duty to obey the law, because laws solve coordination problems. See, for example, Boardman (1987). In the text, I make an analogous point for ground-level moral practices.

4. For this reason, I think Nozick should have chosen another example. It is not plausible that most people would think that their life prospects would be enhanced by forced labor at a radio station, even if they like the music the station plays.

5. Actually, for Hobbes [1651], the consent requirement was not really a moral requirement and, in any case, it was almost vacuous, because any kind of consent, even consent extracted by the threat of death, was adequate.

6. Most, but not all. There may be other ways of specifying a practice to make the practice acceptable. In the Wild Beast example above, the requirement of majority approval of the procedure played this role.

7. In chapter 4, I elaborate my conception of a *participant* to include those who are prevented from participating by incapacity, but who otherwise would be willing to participate.

8. A precise statement of the main principle would require that it be stated as a game theoretic equilibrium rule, in order to resolve the coordination problem discussed in the text (cf. Regan 1980). I ignore that complication here.

9. This is my version of the part of morality that Scanlon refers to as *What We Owe to Each Other* (1998). I should note that in the next chapter I extend the category of nonresponsible noncompliers to those whose inability to cooperate is due to severe mental or physical impairment.

10. Here I am agreeing with Blake (2001 and forthcoming) that what we owe to our fellow citizens is different from what we owe to outsiders. Though

his reasons are nonconsequentialist, I think that the main principle explains why Blake's reasons are morally appropriate as ground-level reasons.

11. G. A. Cohen distinguishes between fundamental principles of justice and rules of regulation for society (2008, 277–278). He argues that Rawls's principles are rules of regulation, not principles of justice. I agree.

12. Nothing crucial hinges on the semantic question of whether the main principle is a consequentialist principle or not. In discussions with philosophical audiences, I have found that there is a roughly equal division between those who think that it is a consequentialist principle and those who think that it is not. Perhaps I should just say that it makes moral improvement more consequentialist than most philosophers have thought.

Chapter 4

1. I could understand why you might choose it if your own life were full of misery with no way out. I hope your life is not like that. Some people would not hook up because they do not think the life of a successful Socrates has very much hedonic value. Then choose another virtual life. Imagine you can have whatever kind of virtual life would maximize net hedonic value. Some people think that the reason most of us would not choose to hook up is that the pain of thinking that we had made the choice would be too great. Suppose we set up the machine so that all we have to do is push a button and it will take over our consciousness. How painful could the few seconds between the decision to push the button and the onset of a virtual life of bliss be? So bad that it could not be outweighed by any amount of pleasure? This is not true. When I offer these hypothetical choices to my students, even though the overwhelming majority of them will not agree to hook up for life, almost all of them will agree to hook up for a few seconds of *believing* they have decided to hook up for life for $10,000 to be spent after they have been unhooked. So it is not true that the experience of believing that they have hooked up for life is so bad that it cannot be outweighed by other good things.

2. Rawls himself seems to have thought that his "thin" theory of the good could serve as an "index of well-being" (1971, 396).

3. Dworkin's (2000) equality of resources account is a complicated case. It avoids the objections I raise here. I discuss it separately in chapter 11.

4. I should note that Sen does not hold that capabilities are the only things that matter in the evaluation of political institutions (1999, 79).

5. Raz also argues against the possibility of separating morality from well-being (1986, 313–320).

6. In saying that the principle evaluates *life* prospects, I mean to deny that the values that make up the value of a life are separable, as they would be if goodness were just a matter of the total sum of pleasure minus pain (and pleasures or pains felt at one time did not affect the intensity of pleasures or pains felt at other times).

7. See, for example, Kahneman and Tversky (1979) and sources cited therein.

8. I myself think that expected utility is not an adequate measure of life prospects, because variance matters. Abstracting away from real-world complications, the issue is easy to explain conceptually. Suppose the utility of a normal life is 100. I don't think that the life prospects of someone with a guaranteed lifetime utility of 100 are equivalent to the life prospects of someone who has a 50–50 chance at 0 or 200.

9. I discuss brute luck and option luck more fully in chapter 11.

10. I say "in most circumstances" because it is always possible to imagine weird scenarios in which, for example, an evil demon threatens to kill everyone who receives the immunization. In such a situation, though the system of forced immunization would reduce the probability of each individual's dying from the fatal disease, it would greatly increase (to one) the probability of their dying at the hands of the alien. This is yet one more example of the defeasibility of moral reasoning.

11. Although the name *original position* is due to Rawls (1971), Harsanyi (1953) proposed a similar thought experiment. It is for this reason that in Talbott (2005), I referred to it as the *Harsanyi-Rawls original position*.

12. Actually, a similar problem arises for Rawls's original position, because similar reasoning could be used by the parties to conclude that they were not young babies (young babies would not have a language to reason with), to give only one example. This is a potentially serious problem for Rawls's original position, because he assumes the parties to be rational and mutually disinterested (1971, 144), and thus they might not be motivated to give consideration to the interests of young babies.

13. It is important to note that I use the term "expectation" in the mathematical sense of an average, not the psychological sense of what someone expects to happen.

14. In later work, Rawls more fully developed his actual-world narrowing of the maximin expectation principle (2001, 66–72 and 97–102). The strategy is always the same, to claim that facts about the actual world exclude the kinds of possibilities that would raise problems for the formula. Even the early Rawls realized that his response was of the same kind as utilitarian arguments that their principles would not favor slavery in this world; so he decided to defend this sort of actual-world narrowing of utilitarianism (1971, 159). This is not the place for an extended discussion of actual-world narrowings, but I should mention that Rawls correctly describes the alternative to his actual-world narrowing as the view that "moral conceptions should hold for all possible worlds" (1971, 159) and then immediately proceeds to caricature the position almost beyond recognition. All that the advocate of this alternative is committed to is the existence of a principle that explains rightness or justice or moral improvement in actual and hypothetical cases. If the principle gave the correct results in all actual and hypothetical cases, it would be true in all possible worlds.

15. It is unfortunate that Rawls never did clearly specify his cutoff date. We can only speculate. No date after birth could plausibly be chosen, because if any time after birth is chosen as the cutoff date, the theory would imply that policies that kill babies or children before that date would raise no issues of justice, so long as they raised the expectations of those who were *not* killed before that date. This could not be right. An alternative that would not require

a cutoff date would be to define the least advantaged group in terms of *actual* income and wealth, rather than in terms of *expectations*. I discuss this alternative shortly. It is clear that Rawls himself is committed to a cutoff date, because he holds that "individuals' expectations of primary goods (their index) can be the same ex ante, while the goods they actually receive are different ex post, depending on the various contingencies . . ." (2001, 173). The cutoff date is the line that separates *ex ante* from *ex post*.

16. Rawls himself acknowledges this when he calls the formula a "maximizing principle" (1971, 79).

17. I first heard the leveling down objection from Nozick, who presents it in general form in (1974, 229, 237), but who presented it in much more vivid form in his course on the book manuscript in the fall of 1972.

18. There is one further defense of Rawls that I only mention here. Perhaps his theory is more ideal than we thought. Perhaps it applies only to beings who cannot become permanently disabled. This would make Rawls's theory even more of an idealization than it already is. I discuss problems with ideal theories in chapter 10.

19. I should note that the later Rawls did try to show that his principles of justice could be applied to at least some issues of health care. In *Justice as Fairness: A Restatement*, he discussed the possibility of applying his two principles to "the medical and health needs of citizens as normal cooperating members of society whose capacities for a time fall below the minimum" (2001, 173), where the minimum is defined as the "essential capacities for being normal and fully cooperating members of a society" (171). The idea is that even if his theory did not apply to those with permanent disabilities, it could be applied to those who were temporarily disabled and needed medical care to be able to be able to return to productive work. However, Rawls is mistaken to think that his theory can even explain why such medical care would be *required* by justice. He is correct to argue that his principles *permit* ex post differences in primary goods "depending on illnesses and accidents" (2001, 173), but the problem is that his principles don't *require* any such differences. His principles *require* maximizing only the ex ante average (the index). They place no constraints on the distribution of primary goods ex post. Therefore, they could be satisfied by a system that did not provide any guarantee of health care, not even care that would enable temporarily disabled workers to return to work. I should note that there is a way that some kind of right to medical care could be justified in Rawls's theory. Consider only those types of medical care that tend to make workers more productive. Call that *productive medical care*. The difference principle could easily justify a right to productive medical care, if the resulting increase in productivity raised the expectations of the LAG.

20. Analytically, it is easier to define the *noncooperators* first as the responsible noncompliers (when the responsible noncomplier exclusion obtains) and then to define *cooperators* as all potential cooperators who are not noncooperators.

21. Of course, the explanation of the subsequent worldwide collapse of Marxist economies is that Marx's theory about what made capitalism inequitable (viz., private ownership of the means of production) was a big mistake. I discuss this more fully in chapter 9.

22. See Talbott (2005, chap. 8) for a fuller discussion of these issues.

Chapter 5

1. There is a second aspect to the metaphysical mystery. It is to explain why normative truths are motivating. This is a mystery for all normative truths, including truths about what it is rational to believe and what it is rational to do in nonmoral contexts. To solve this part of the mystery, I follow Korsgaard (1986), though I use her strategy to defend a kind of substantive normative realism that she rejects (1996). Her proposal is that although normative truths are not necessarily motivating for human beings, that is only because human beings are not necessarily rational. Normative truths are motivating to rational (or reasonable) agents, because being rational (or reasonable) involves being *responsive* to reasons. (The term comes from Nozick 1993.) Consider the example of the person who believes that p and recognizes that p implies q, but does not believe q (and does not stop believing p). To be rational it is not enough to be *sensitive* to reasons—for example, to recognize that p implies q. It also requires some level of *responsiveness* to reasons—in this case, either to accept q or give up p, as appropriate. A similar responsiveness is a component of nonmoral practical rationality. Someone who was sensitive to the requirement of transitivity of preferences would be able to recognize that her ranking of alternatives A, B, and C was intransitive. If she were not motivated to change the ranking to a transitive one, she would fail to be appropriately responsive to her recognition of the intransitivity and thereby would exhibit a failure of rationality on her part (cf. Hampton 1992). Finally, following Rawls (1993), we can say that to be reasonable, one must be sensitive to and responsive to reasonableness, understood as cooperating on fair terms of social cooperation, where being sensitive to and responsive to reasonableness is possible without thinking of one's acts in Rawls's terms (e.g., thinking of oneself as cooperating or not cooperating on fair terms of social cooperation). Sensitivity and responsiveness would both be explained subjunctively. There are many grades of sensitivity and responsiveness. Consider a particularly simple example. Sensitivity to reasonableness might be captured subjunctively by the fact that if an act was not cooperating on fair terms of social cooperation, one would generally not think it was right. Responsiveness to reasonableness might be captured subjunctively by the fact that if one did not think an act was right, one would generally not do it. As I discuss shortly, the person of practical wisdom would be someone who is both sensitive to and responsive to reasonableness. See Railton (1984) for a similar subjunctive account of responsiveness to the utilitarian maximizing formula.

2. Nozick gives a name to this subjunctive sensitivity condition. It is a necessary condition for a true belief to *track the truth* (1981, chap. 3). Others had proposed subjunctive conditions of this kind for knowledge even before Nozick—for example, Dretske (1971) and, for perceptual knowledge, Goldman (1976).

3. Dworkin (1996) simply claims that, even if we don't understand how we do it, it makes more sense for us to believe that we are able to make reliable moral judgments than to think that there is nothing wrong with slavery or genocide. I agree. But it would be even better if we could understand how we are able to make reliable moral judgments (at least in clear cases).

4. It goes without saying that a failed experiment does not refute a principle. Here I am describing a historical development, not making a philosophical argument. Also, direct consequentialists will challenge the historical accuracy of my claim that the Marxist dictators were actually applying a consequentialist principle. I believe they may be correct. However, I still think that the Marxist dictatorships discredited direct consequentialism, because those dictators did intend the kinds of changes—for example, increasing gross domestic product—that a direct consequentialist would favor and their attempts to do so were utter failures. China's economy began significant growth only when the nominally communist government replaced Marxist economics with private property rights, rights that operate as constraints on direct government action. I discuss property rights in chapter 9.

5. *West Coast Hotel Co. v. Parrish*, 300 U.S. 379 (1937).

6. Of course, this is no refutation of Kantianism or libertarianism. However, it represents the historical discovery of a philosophical problem for both kinds of view, which is their failure to allow for governments to enact coercive solutions to collective action problems without obtaining unanimous consent from their citizens (which, of course, could almost never be obtained).

7. Though China's government labels itself *communist*, it ceased to be so when it began to recognize property rights. This is another reminder that names need not be descriptions of what they name.

8. Not only Plato, but even today lots of intellectuals favor advocate replacing jury trials with judge trials. Thus, the following report is quite striking: In a survey of federal judges "97 percent of the 594 federal judges surveyed said they agree with the jury verdicts most or all of the time. By an 8-to-1 ratio, federal judges said that if they were on trial, they would prefer to have their dispute decided by a jury rather than a judge" (Curriden 2000, 52). See also one federal judge's defense of the jury system (Dwyer 2004).

9. It is important not to overstate the disadvantages of having inconsistent beliefs. There is good reason to believe that almost everyone has some inconsistent beliefs. We could spend all our waking hours ferreting out inconsistencies. This would not be an evolutionarily advantageous way of spending our waking hours. Although it would be counterproductive to spend every waking hour looking for inconsistencies, it would also be counterproductive to ignore them when we find them.

10. If the law of non-contradiction is a ground-level principle of rational belief, we should not be surprised to find that it has exceptions (e.g., Priest 2006). This leads me to think that there is a meta-level principle that explains these exceptions, also, but the question of the existence of normative truths about what it is rational to believe is obviously a larger topic than I can address here.

11. In this note I provide a more precise specification of the difference between explicitly applying a moral rule or principle and being implicitly sensitive to one, on my consequentialist account. I suppose that there is a complex probabilistic function that assigns a value for *goodness expectancy* (GE) to an act based on information about the act and the practice of which it is a part, and the implementation practices if it involves a change to existing moral practices. (Think of goodness expectancy on analogy with life

expectancy.) Then consider the situation of George deciding between two acts, to lie (L) or to tell the truth (–L). George is an explicit rule follower. He reasons as follows: Lying is wrong, therefore I ought not to lie. At the meta-level, this reasoning is modeled by the following probabilistic inequality:

(1) GE(-L) > GE(L).

Because GE is a probabilistic function, it has a reference class logic. Adding more information about the situation—for example, that by lying one could save the life of an innocent person, the inequality reverses. Let S = Save the life of an innocent person):

(3) GE(L&S) > GE(-L&-S).

This is how explicit rule or principle following would be modeled at the meta-theoretical level. The rule against lying would have an explicit exception for lies that would save the life of an innocent person. How can we model the reasoning of the person of practical wisdom, if the person of practical wisdom does not explicitly apply rules or principles? Consider the person of practical wisdom Jane, who is deciding between two acts, A and -A. Jane does not explicitly apply moral rules or principles. She simply responds to the situation, emotionally as well as cognitively, and comes to a decision about the right thing to do—for example, that act A is the right thing to do. At the meta-level, Jane's implicit sensitivity to GE is modeled by the following inequality:

(2) GE(A/Jane believes that A is the right thing to do) > GE(-A/Jane believes that A is the right thing to do).

Jane does not calculate GE. She does not even have to classify the act under an explicit rule or principle (e.g., as lying or telling the truth). Undoubtedly, she does classify the act in many categories. But her sensitivity can go beyond any of the categories she uses to classify the act. She can just have a feeling that -A would turn out badly. To be a reliable classifier of right and wrong acts, she does not have to do any explicit reasoning at all. Her cognitive processing just has to be sensitive to the factors that determine the value of the probabilistic function GE.

Note that Jane could not explicitly apply the inequality (2) in her reasoning about what to do, but someone else could. If I have a choice between A and -A, I can ask Jane which one would be the right choice. When she tells me that it is A, I can explicitly apply inequality (2) to decide what to do. However, moral training would not work unless Jane could teach others to dispense with inequalities like (2) and to develop their own sensitivity to GE.

12. For example, in *Riggs v. Palmer*, 115 N.Y. 506, 22 N.E. 188 (1889) (holding that a murderer cannot recover under the will of the person he murdered) or in *Henningsen v. Bloomfield Motors, Inc.*, 32 N.J. 358, 161 A.2d 69 (1960) (establishing strict liability of manufactures for injuries due to a defective automobile, not dependent on privity of contract).

13. 32 N.J. 358, 161 A.2d 69 (1960).

14. There is a second source of apparent unfairness in the *Henningsen* case. Because the court adopted a standard of strict liability for damages, it would seem that it unfairly treated nonresponsible noncompliers—those manufacturers who were not at fault for the fact that their products were unsafe. However, the main principle could endorse a system of strict civil

liability if it equitably promoted the life prospects of compliers and nonresponsible noncompliers. I explain how such a system could equitably promote their life prospects in chapter 9. Even more surprisingly, in the next chapter, I show how, at least in theory, a system of strict *criminal* liability could be endorsed by the main principle as equitably promoting the life prospects of compliers and nonresponsible noncompliers.

15. I am grateful to Liam Murphy for suggesting to me that Dworkin's account of legal interpretation might have been motivated, in part, to solve this problem of retroactivity.

Chapter 6

1. Even in the twenty-first century, kin deterrence solutions to security CAPs are still in effect in some parts of the world. For example, in Albania, the code of revenge known as *Kanun* requires that a killing of a family member be avenged by killing a male relative of the killer. In May 2002, Albanian Isa Haruni and his male relatives lived in fear that they could be killed any day, because of a feud sparked by a killing committed by his cousin seven years earlier. Once begun, the feuds are potentially endless. See Dhimgjoka (2002).

2. My claim here is that, when a government satisfies the main principle, it would not be justifiable self-defense for a suspect who was reasonably, though mistakenly, thought to be guilty, to kill the sheriff who came to arrest her. Indeed, I believe it would be wrong to kill a jail guard to escape the death penalty for a murder that one did not commit. In this I disagree with Hobbes [1651], chap. 14. But I do not go as far as Socrates, who seemed to hold that if one were wrongly condemned to death, one would have a duty not to do anything to prevent the sentence from being carried out, and that it would be wrong to escape, even if one could do so without directly harming anyone (Plato, *Crito*, 50a6–54e2).

3. Because I have not claimed to provide a theory of punishment, the main principle may not be the sole source of prisoners' rights. But some prisoners' rights would be justified by the main principle because of the possibility of convicting an innocent person.

4. 384 U.S. 436 (1966).

5. *Mapp v. Ohio*, *367 U.S. 643* (1961).

6. *Johnson v. New Jersey*, 384 U.S. 719 (1966).

7. *Griffith v. Kentucky*, 479 U.S. 314 (1987).

8. Kant famously held that punishment is not optional, but required by the moral law and that even the sovereign's right to grant clemency is limited to offenses committed against the sovereign. According to Kant, the duty to punish creates an absolute moral obligation to kill all convicted murderers and no exceptions can be justified on consequentialist grounds ([1797], 104–110).

9. Many people, even Supreme Court justices, are in denial about this. In *Kansas v. Marsh* Justice Scalia endorsed the wrongful conviction rate calculated by Joshua Marquis (2006) of .027—that is, a success rate of 99.973%. Marquis made his calculation by dividing the number of exonerations due to

DNA evidence by the total number of felony convictions in the United States. On Marquis's methodology, we could increase our success rate by refusing DNA tests for convicted defendants. Indeed, using Marquis's methodology, China, which convicts 99% of its defendants, could claim a 100% success rate, because almost none of them are ever exonerated.

10. These rates are surely too high. There are many reforms that would reduce the incidence of wrongful conviction, including videotaping all police interrogations, replacing standard police lineups with sequential lineups, and not allowing a conviction to be based on eyewitness identification alone or based on plea bargained testimony alone. Prosecutors have opposed all these measures.

11. The discussion in the text greatly understates the problems with the existing criminal justice system in the United States. In the United States prosecutors often obtain a conviction when, by any measure, the objective probability of guilt is much less than .5. Here is a brief outline of how it happens. To solve a crime, the police and prosecutors try to determine, on the basis of the available evidence, who the most probable perpetrator is. When there is good evidence, the probability that the most probable perpetrator committed the crime may be 90% or higher. When there is little evidence and few leads, anyone who happens to have been in the vicinity may turn out to be the most probable perpetrator, even though the probability that s/he is the perpetrator is 10% or less. Nonetheless, once the police and prosecutors fix on the most probable perpetrator, in many jurisdictions there is a high probability that they will put together a strong enough case to get to trial and a high probability that they will obtain a conviction. Thus, someone who just happens to be in the wrong neighborhood at the wrong time or who is unfortunate enough not to have a witness to support an alibi (or is unfortunate enough that the witness who supports his alibi dies before trial) is likely to be convicted even though, by any reasonable objective measure, the probability that he committed the crime is extremely low, much lower than .5. See Grisham (2006) for an extended description of how this process works. Grisham tells the story of two defendants who were mistakenly convicted of murder, one of whom spent time on death row. They were eventually exonerated by DNA evidence. However, in telling their story, he also recounts the convictions of two other defendants who are still in prison because there was no DNA evidence and thus no exoneration for them, even though it is clear that they are almost certainly innocent. I confine this discussion to a note, because I take it to be evident that no such criminal justice practices would be endorsed by the main principle or by any plausible nonconsequentialist view. For the most thorough review of the types of evidence that led to wrongful convictions of defendants later exonerated by DNA evidence, see Garrett (2008).

12. A related issue is the question of the jury vote required for a criminal conviction. Not every jurisdiction requires unanimity for criminal verdicts. Is unanimity a moral requirement? Is there a nonconsequentialist consideration that helps to decide what this requirement should be? Of course, for the consequentialist, the considerations will be the same as the general considerations for balancing mistaken convictions against deterrence.

13. There is no agreed upon statement of the doctrine of double effect. See Woodward (2001) for a good collection of defenses and criticisms of

versions of the doctrine. For a criticism of the distinction between intending and foreseeing and variants on it, see Mark Johnston's proposed counterexample reported by Delaney (2007).

14. Not all defenders of DDE are absolutists. Quinn (1989), for example, defends a nonabsolutist version of DDE. Because I think some kind of nonabsolutist version of DDE would be favored by the main principle, I do not regard such views as necessarily nonconsequentialist.

15. Suppose we had a practice of permitting prosecutors to cook the evidence against a defendant in circumstances C. There would be two sources of abuse. First, there would be prosecutors who would intentionally flout the standard and cook the evidence even when circumstances C did not hold. Second, there would be prosecutors who would deceive themselves into believing that circumstances C held when they did not, in order to feel justified in cooking the evidence. I would not be surprised if the latter kind of abuse were even more significant than the former, though they would both be serious. By the way, this also explains why any practice of making exceptions to a torture prohibition would also generate lots of abuse and why, therefore, the main principle would not endorse any practice of making exceptions to a torture prohibition (cf. Mayerfeld 2008).

16. There are actually a number of examples of strict liability (i.e., liability not based on a determination of fault) in the criminal law, including statutory rape, selling alcoholic beverages to minors, selling adulterated milk, and many traffic laws (e.g., speeding laws). Also, there is the legal principle that ignorance of the law is no excuse.

17. Actually, although Thomson clearly has sympathies for some sort of hypothetical consent as a test for legitimacy (even if not a full explanation of it), she does not actually commit herself to going beyond a libertarian account of government legitimacy (1990, 361).

18. In July 2006, U.N. Secretary-General Annan released a report that identified violence against women as a human rights violation.

Chapter 7

1. The complete history would begin with Hegel, for it was Hegel's [1821] theory of history as the revelation of absolute spirit that first located rationality in a historical process. However, Hegel did not make the crucial move of democratizing the process.

2. I address in this note the question of whether Mill's epistemology can apply to logic and mathematics, which seem to be areas of inquiry that do employ *a priori* justification. Developments in both these areas since Mill have provided support for his epistemology. In logic, Frege thought that his axioms for set theory were justified *a priori*, until he received Russell's letter showing that they were inconsistent. Even after Frege's painful experience, other attempts to axiomatize set theory were discovered to be inconsistent. It was later proved that any proof of the consistency of a mathematical system adequate to express arithmetic would have to be given in a stronger (and thus more likely to be inconsistent) formal system. Similarly, Andrew Weil's initial "proof" of Fermat's Last Theorem was found by a colleague to be flawed.

Weil fixed the flaw and published a new proof. Does the new proof have a flaw? Well, no one has found one. So the best evidence of consistency of mathematical systems or of the validity of mathematical proofs is that no one has found an inconsistency or a flaw. Mill's point is that the lack of contrary evidence can provide rational support only if people are free to challenge the consistency of a mathematical system or the validity of a proof. Thus, even logic and mathematics get their rational support from a social process of the free give-and-take of opinion.

3. This is not to say that there were no philosophers who advocated transcendent normative truths in either tradition. It is only to try to characterize the dominant view.

4. I quote his revised statement of the two principles in note 9 to chapter 1.

5. Here is how he characterized a liberal political conception of justice:

> The content of such a conception is given by three main features: first, a specification of certain basic rights, liberties and opportunities (of a kind familiar from constitutional democratic regimes); second, an assignment of special priority to those rights, liberties, and opportunities, especially with respect to claims of the general good and of perfectionist values; and third, measures assuring to all citizens adequate all-purpose means to make effective use of their liberties and opportunities. These elements can be understood in different ways, so that there are many variant liberalisms. (1993, 6)

6. Although the Nazism example is more attention-getting, it is not necessary to look outside of Rawls's own work to find a reasonable disagreement on the nature of reasonableness. As I have already said, as *reasonable* is used in *Political Liberalism*, it is unreasonable not to agree on all the basic rights guaranteed by Rawls's first principle of justice. Call these the *liberal rights*. However, in *The Law of Peoples*, when Rawls considered the question of which rights are universal human rights, his answer was: only a small subset of the liberal rights in *Political Liberalism*. By arguing in *The Law of Peoples* that not all liberal rights are human rights, he is implicitly conceding that there can be reasonable disagreement on liberal rights, or, in the terms introduced earlier, it can be reasonable to be unreasonable$_L$. Again, Rawls's political liberalism turns out to be a kind of moral relativism.

7. Note that it could be rational to give up this presupposition even if the voice in the bush was not the voice of God and what it had told you about the laws of the universe and about what happened before the Big Bang were not true. The issue here is simply whether the presupposition that identifies truth or purely descriptive validity with the results of the ideal process of discourse is inescapable.

8. There are many theories of self-deception that do not require that the person who is self-deceived in believing that p also believe–p. See, for example, Mele (2001) and Talbott (1995).

9. Unless the presupposition were stated generally as: I am committed to my statement surviving in an ideal process governed by the norms most conducive to determining the truth. I think there might be an attenuated sense in

which we are committed to something like this when we make a statement, but Habermas could not endorse this position, because it would require that normative validity be aimed at normative truth. In my opinion, perhaps the most powerful evidence that normative validity is aimed at truth is that exactly the same *process* of discourse is determinative of both purely descriptive and normative validity. Also, why would truth-preserving rules of logic be useful in a domain that has nothing to do with truth? Habermas offers an alternative explanation of the usefulness of deductive logic and other forms of reasoning in the normative realm (2003, 266–271). Gibbard (1990) offers a noncognitivist proposal for solving this problem.

10. A philosophy professor should be especially aware of the potential for a process of argumentation to generate division, not consensus. The level of argumentation in philosophy department deliberations is quite high. However, even when the issue concerns the common good and everyone sincerely aims at the common good, for example, in a hiring decision, argumentation can produce, not consensus, but bitter and unbridgeable divisions.

Chapter 8

1. Mill would approve of the higher standard of libel for public figures announced by the U.S. Supreme Court in *New York Times v. Sullivan*, 376 U.S. 254 (1964). In that case, the court limited libel actions by public figures to cases in which the defendant showed actual malice—that is, that the defendant knew his statement to be false or made it in reckless disregard for the truth. There are still many places in the world where to criticize a public official is to risk being sued for libel. Typically, the decision concerning whether the critical statement was true will be made by a judge appointed by the plaintiff.

2. 341 U.S. 494 (1951). I discuss the *Dennis* case at length shortly.

3. *Brandenburg v. Ohio*, 395 U.S. 444 (1969).

4. "Reasonable disagreement is disagreement between reasonable persons" (1993, 55; also 39).

5. Rawls does not specifically say that views that reject other parts of the liberal conception of justice are unreasonable, but it would seem he would be committed to that result, at least if the rejection is not trivial, because it is the entire liberal conception that specifies the fair terms of social cooperation in Rawls's theory. Nothing crucial hinges on this, because my discussion will focus on the rights in the first principle.

6. When he wrote the preface to the paperback edition of *Political Liberalism*, Rawls tried to soften the impression that his standard was intolerant of reasonable disagreement with liberalism, by allowing that there is a family of liberal conceptions that are acceptable (1994, lii–liii). This softening does not conflict with anything that I say in the text about Rawls.

7. Freeman (2001) has argued persuasively that libertarianism is not a liberal view. He did not draw the obvious conclusion that libertarianism is not reasonable$_L$, but he shows in a different article that he has the premises to do so (2000, 411).

8. This is a somewhat startling result, given Rawls's attitude toward unreasonable views. He says that they seem to be permanent fact of life, which "gives us the practical task of containing them—like war and disease—so that they do not overturn political justice" (1993, 64 n. 19).

9. There is a potential for the same kind of ambiguity to infect moral philosophy. For example, Scanlon's (1998) account of morality in terms of reasons that no one could reasonably reject avoids this ambiguity in "reasonably" only if he doesn't try to use his formula to try to resolve any reasonable (in the ordinary sense) moral disagreements.

10. For my criticism of his list, see Talbott (2005, 10–13).

11. Even though I made no use of the reasonable disagreement test in my account of human rights in the first volume (Talbott 2005), some reviewers just assumed that I must be committed to a no-reasonable-disagreement criterion of human rights. See, for example, Reidy (2008) and von Platz (2008) and my replies (Talbott 2008). I herewith disclaim any use of the distinction in my theory of human rights and acknowledge that there is reasonable disagreement on just about everything I said in the first volume and just about everything I say in this volume. I now believe that I should have included the acknowledgment of reasonable disagreement in my conception of epistemic modesty (cf. Talbott 2005, 15).

12. This point has been made by a number of authors, including Gaus (1997), Christiano (1997), and Waldron (1998).

13. For a careful and largely critical investigation of the role of reasonable agreement and overlapping consensus in a theory of human rights, see Kim (2009). Kim suggests that agreement plays no role in the conception of a human right, but it can play a role in the justification of the use of coercion to enforce it. I think it is important to distinguish between forcible intervention that is paternalistic toward adults and forcible intervention that is not. When the intervention is not paternalistic—for example, intervention to prevent mass rape, I see no reason to think it cannot be done in a way that would be endorsed by the main principle. In such a case, the intervention is not paternalistic, because it does not overrule the judgment of those it aims to assist. Those who are potential and actual victims of rape want it to stop. In contrast, I do not think that the main principle endorses paternalistic intervention—for example, intervention to force human rights on a resistant population for their own good. Persuasion is a much preferred implementation practice.

14. If one believes, as I do, that the majority opinion of experts in any area is more likely to be true than a minority opinion, how could it be rational for any expert to stick with an opinion he knows to be a minority opinion? See Kitcher (1990) and Pettit (2006) for attempts to answer this question. I think that a complete answer requires us to explain individual rationality by reference to the role of minority opinions in moving majority opinion closer to the truth over time.

15. *Citizens United v. Federal Election Commission*, 558 U.S. 50 (2010).

16. In addition, 40% of scientists believed that human beings evolved from lower forms of life in a process guided by God. Only 5% of scientists (as opposed to 44% of the general population) denied that human beings evolved from less advanced forms of life.

17. It is worth mentioning that there will always be examples of the opposite kind, also. For example, during the 1960s, in the heyday of behaviorism in academic psychology, a far higher percentage of academic psychologists than the general public would have denied the existence of mental states. I believe the commonsense view was closer to the truth. An even more striking example is afforded by epistemologists, purported experts in what we know or are epistemically justified in believing. For most of the past 400 years, a majority of Western epistemologists would have denied that we can have any knowledge of external objects. Again, I think that the commonsense view has been closer to the truth. Someone who does not believe that there are mental states or that we have any knowledge of external objects will not be swayed by these examples.

18. For a much fuller discussion of the Web's potential for aggregating opinions, see Sunstein (2006b).

19. Sunstein (2006a) rightly criticizes juries and other deliberative bodies as less than optimal information sharing institutions, because they tend to reinforce majority opinion and extinguish minority opinion. In eliciting minority information, as he argues, prediction markets are superior to deliberative bodies. This is a case of whether you see the glass as half full or half empty. It is the very tendency of the jury system to reinforce majority opinion that is its greatest epistemic virtue in criminal trials. But it is far from epistemically ideal.

20. For an example of a very modest epistemic defense of democracy, see Estlund (2008). For a more ambitious epistemic defense of liberal institutions and human rights, very much along the lines of the position I articulate here, see Buchanan (2004b and 2008).

21. A generalization because Mill had a hedonistic conception of well-being and I believe that even the nature of well-being itself is something that needs to be discovered.

22. It is often forgotten what a risky business challenging authority used to be. We are all familiar with the threats that forced Galileo to recant any claim to truth. We sometimes forget that many of the most important works of philosophy were published under pseudonyms, for fear of persecution. And many philosophers, including Hobbes, Locke, and Rousseau, have, at one time or another, had to flee for their lives to avoid persecution.

23. The argument also supports some sort of market-based economic system and some sort of right against paternalism, but I defer the discussion of economic rights to the next chapter and of rights against paternalism to chapters 12 and 13.

24. 341 U.S. 494 (1951).

25. *Masses Publishing Co. v. Patten* 245 F. 535 (1917).

26. 395 U.S. 444 (1969).

27. For more on the right to education, see Gutmann (1987).

Chapter 9

1. In invoking Rousseau's notion of "conventions" I mean to block the inference made by some social constructionists that moral constraints are

themselves social constructions. Rousseau's conventions were subject to a powerful moral constraint (that they express the general will). Similarly, on my view, property rights are social constructions that are morally constrained by the main principle, which is not itself a social construction.

2. It comes as something of a surprise to find out that, even without legal institutions, California gold miners relied on a package of shared conventions for deciding ownership. When a gold strike was made, the word spread quickly and other miners would arrive at the camp. Rather than a free-for-all to determine ownership, the group as a whole would mark off claims of a size that one or two men could work, and then miners would choose their claims in the order they had arrived at the camp. It is often thought that the gold fields were violent, because they were literally lawless. However, they were generally nonviolent and personal property was surprisingly secure (Zerbe and Anderson 2001).

3. Also, there are some advantages to the delay. In the current system, patients in the wealthier countries are the guinea pigs whose use of new drugs makes possible a more reliable determination of their safety hazards.

4. Symmetry considerations suggest the possibility that ownership rights in tangible property should also expire. However, for the social constructionist there are important differences between tangible and intangible property that explain why only rights to the latter should expire. Consider, for example, what would happen if, as suggested in the Old Testament (Leviticus 25: 10–13), every 50 years were a jubilee year and all property rights to real property expired so that it could be redistributed. As the jubilee year approached, no owners would be motivated to invest in improvements to their property and, even if they were, no lenders would be motivated to loan money for improvements. As the jubilee year approached, owners would become more like renters and the value of real property would inexorably decline. Of course, intellectual property needs no maintenance, so there are no such costs to allowing intellectual property rights to expire.

5. Of course, even if both parties to an economic transaction increase their life prospects, the transaction may have externalities that reduce other people's life prospects. This is just a reminder that economic rights must be embedded in a framework of other rights.

6. As an empirical matter, Hobbes was mistaken. There have been many stable systems of property rights, including common property rights, that were based on group enforcement (as in a state of nature) with no need for Hobbes's sovereign.

7. This is an instance of Hegel's [1821] famous saying that the Owl of Minerva flies at dusk. Hegel was speaking of philosophers, but the saying applies to everyone who makes recommendations for political change. As a species, we are much better equipped to evaluate a change *ex post* than to predict its results *ex ante*. The rationale for democratic rights crucially depends on this fact, as I explain in chapter 10.

8. It is true that a majority of the world's population cannot afford to buy a personal computer or a cell phone or a sewing machine, but the proportion who can afford them is growing dramatically.

9. Of course, the real world is much more complicated than this simple model. Someone with an idea for a new product would have to attract

financing. Still, those who come up with the ideas and those who provide the financing are part of a system that rewards both the inventors and funders of successful ideas.

10. As selection processes, markets differ from processes such as natural selection in an important way. Market systems motivate innovation. In natural selection, mutations occur randomly; in a market system, innovations are not random. Entrepreneurs are attempting to improve on the status quo. So a market system might better be described as an incentivized selection process.

11. This assumes that aggregate willingness to pay correlates positively with increases in life prospects. Of course, what it correlates with is an increase in the life prospects of those with money. This complication will be addressed shortly.

12. There was a time when water's commodity value was also zero, or close to it. Not any longer. Are we approaching the end of free air?

13. Although I take these examples, by and large, to be uncontroversial examples of improvements endorsed by the main principle, that they are improvements is a function of the empirical consequences of adopting them as policies. Although the reasons given by the courts for making these changes have often been stated in terms of fairness, I believe that the applicability of those very standards of fairness is explained by the main principle.

14. In 2010 in the United States most commercial sales are governed by the Uniform Commercial Code, first promulgated in 1952 and ultimately adopted by all 50 states. However, that code is itself the culmination of centuries of developments in the common law. The relevance of the UCC here is that it is largely a compendium of the qualifications and exceptions that have been developed in the common law to the actual consent exception to libertarian natural rights—that is, to the simple rule that parties are bound by their voluntary agreements and that a party is entitled to damages for its reasonable losses caused by another party's breach of such an agreement.

15. The doctrine of implied warranty was reformulated as a doctrine of strict liability in *Greenman v. Yuba Power Products, Inc.*, 59 C2d 57 (1963).

16. 32 N.J. 358, 161 A.2d 69 (1960).

17. The requirement of privity for recovery based on negligence was eliminated in *MacPherson v. Buick Motor Co.*, 217 N.Y. 382, 111 N.E. 1050 (1916).

18. First enunciated in the British case *Rylands v. Fletcher*, LR 3 HL 330 (1868). The doctrine of strict liability has the effect of spreading the costs of compensation for injuries over the entire population of buyers, those who are uninjured and those who are injured. I discuss this cost-spreading effect in chapter 11.

19. The classic work on liability for accidents is Calabresi (1970). Calabresi was describing the evolution of standards of liability in terms that can easily be seen to fit the main principle. Calabresi points out that even allowing people to insure for the costs of accidents is a move away from traditional fault-based liability. Traditional notions of liability based on fault transferred the costs of accidents from the victim to the responsible party. Insurance has the effect of spreading the accident costs over a larger population of everyone who buys insurance (i.e., including those who cause accidents and those

who do not), rather than concentrating the costs on those who cause the accidents. This is the first step away from traditional fault-based liability.

My proposal in the text is that the main principle favors a system of liability that has the effect of selecting for activities that can "pay their own way" and against activities that do not. I should emphasize that whether or not standards of strict liability for product defects is such a system of liability is an empirical question.

In chapter 6, I asked the following question: How can a nonconsequentialist justify a criminal justice system that punishes the innocent? A corresponding question here is this: How can a nonconsequentialist justify any system of tort liability that places a dollar value on human life? And yet there could not be an action for wrongful death without some way of making such a determination.

20. Here I am referring to the cost-spreading effects of the doctrine of strict liability. As I explain in chapter 11, cost spreading is aimed at promoting equity, and this kind of reason is not paternalistic. I should also mention that strict liability is found in many areas of noncriminal law. For example, workers' compensation statutes provide compensation to workers for on-the-job injuries even if no one was at fault.

21. I happen to live in Seattle, which is one of the centers of this development, which was begun many years ago by REI (Recreational Equipment, Inc.), when it adopted its policy of accepting returns on items that are no longer serviceable *with no time limit* and then extended by Nordstrom when it adopted its policy of accepting returns for any reason.

22. Most contracts are win-win contracts. The most important category of win-lose contracts are investment purchases. If I buy stock from you and it goes up, you would be motivated to cancel the contract *ex post*; if it goes down, I would be motivated to cancel it *ex post*. So there is no way to replace *ex ante* agreement with *ex post* agreement in these kinds of contracts. Nonetheless, there have been many legal changes to reduce the problem of information asymmetries (e.g., mandatory disclosure requirements for prospectuses and prohibitions on insider trading). Understood nonconsequentially, prohibitions on insider trading look like punishments for doing something wrong. From my consequentialist point of view, they are constraints that attempt to solve the problem of information asymmetries, not because trading on the basis of information asymmetries is inherently wrong, but because allowing insiders to benefit from their knowledge would raise the probability of one-sided trades and thus reduce the investment returns to outsiders and hence reduce their willingness to invest. Thus, prohibiting insider trading is probably a solution to a CAP for insiders. It is in the interest of each insider to trade on his insider information, but it may be worse for insiders as a group if the result is to drive noninsiders from the market.

23. American Law Institute (1981, sections 7–8).

24. In a brief discussion of strict liability, Scanlon indicates that his own nonlibertarian view would draw a sharp line between liability based on fault and strict liability. Because he addresses this issue as part of a discussion of punishment, it is possible that he means to be addressing only the issue of criminal penalties. And, in the end, he says only that laws establishing

penalties for harms based on strict liability would be "more difficult to justify" (1998, 266), thus not completely ruling out that they could be justified.

25. Scanlon actually discusses two kinds of noninstrumental value, *representative* and *symbolic* value, but denies that they are mutually exclusive or exhaustive of noninstrumental value. His argument depends on there being only some noninstrumental value (1998, 253), so I simplify the discussion by focusing on that.

26. Of course, it would always be possible to hold that the value of long-term committed relationships is simply a product of the strength of will they exhibit, in which case long-term relationships in which both parties were miserable would have the most value, because they would require the most willpower to maintain. If developing such willpower had other indirect positive effects on well-being, then perhaps such relationships could be given an indirect consequentialist defense. But if such relationships made the parties miserable and their children miserable, there would be no plausible consequentialist defense of them. Is there some other kind of defense?

For a more extended argument against the view that commitment is itself of intrinsic value, see Calhoun (2009). Calhoun does not defend a consequentialist position, but her account can be easily accommodated within my indirect consequentialist framework.

27. The manufacturer, Peanut Corporate of America, went bankrupt. This is an example of how markets operate as selection processes. But it would be a mistake to argue, as the *Wall Street Journal* editorial board did, that this kind of market discipline is a good substitute for government regulation in cases involving the potential for serious harm. In a market system, it is rational for a manufacturer facing bankruptcy to do anything necessary to maintain solvency (cut back on health and safety expenditures, raid the employee pension fund), unless there are nonmarket sanctions for doing so, because the possibility of bankruptcy is not an effective disincentive for someone who is already facing probable bankruptcy. In the absence of government regulations, it is inevitable that there will be sellers who will cut corners for a competitive advantage, when the alternative is probable bankruptcy. Though not inevitable, there will also usually be some sellers willing to risk the possibility of future bankruptcy for larger current profits.

28. It is surprising that Mill did not recognize the logic of a CAP in prohibitions on slavery contracts (though, of course, he did not have the term), because he did recognize the logic in his discussions of other kinds of labor laws ([1859], 102). Once the logic is clear, it makes sense to say that prohibitions on slavery and indentured servitude are the first minimum wage laws—the minimum minimum wage laws.

29. It is controversial whether minimum wage laws do increase unemployment. See Card and Krueger (1994).

30. The negative income tax was proposed by Friedman (1962, chap. 12), as a way of solving the equity problem without undermining efficiency. As I use the term, it is a way of increasing the earnings of workers (e.g., an earned income credit), not a way of funding a basic income for all, workers and nonworkers alike. I discuss a basic income in chapter 11.

31. In the text I focus on the incentives to the high-wage earners. Of course, the negative income tax would also have to be phased out in a

gradual way to preserve the income ranking of low-wage jobs. For example, it would not make sense to define the negative income tax in such a way that workers in minimum wage jobs would have higher net earnings than workers in jobs with higher pay rates.

32. Reed Hastings (2009), the chief executive of Netflix, has suggested that the marginal tax rate on people with his income be raised to 50%.

33. For example, the gap between those with a college education and those with only a high school education has been increasing for decades. And those without a high school education have substantially lower lifetime earnings than high school graduates. See Bosworth, Burtless, and Sahm (2001).

34. Actually, in Dworkin (2000) we can track the evolution in his thought from evaluating equality over the course of an entire life (chapter 2) to evaluating it at a single point, *ex ante* (chapter 9). I discuss Dworkin's view more fully in chapter 11.

35. Qualified, because labor unions also generate CAPs, for example, by inviting corruption that results in the union leaders sacrificing the good of the workers for their own advantage.

36. Most theorists at least agree that unconscionability doctrine is a paternalistic doctrine. I agree with Shiffrin (2000) that this is a mistake, but I diagnose the mistake differently. In my diagnosis, unconscionability doctrine is a solution to a CAP.

37. I should mention that the main principle endorses solving some capitalists' CAPs—for example, many pollution problems or overuse of resources problems are capitalists' CAPs.

38. An example of another change in traditional norms endorsed by the main principle is the evolution of norms against usury. Traditional Catholicism banned usury, understood as charging interest on loans. It is still prohibited in Islam. In the West, the doctrine has evolved to ban charging excessive interest. So it has evolved into a kind of unconscionability constraint. Bans on charging interest reduce life prospects because, just as paying workers their commodity value motivates them to their most productive activity, allowing borrowers to charge interest on loans directs investments to their most productive use, and thus tends to promote life prospects.

39. In moral terms, the problem with the K-H criterion is that it leaves totally unexplained how the fact that one party to a potential exchange (the winner) would gain more than the other (the loser) could justify conferring on the winner all of the potential gains of the exchange without subtracting any of the potential costs (i.e., without providing any of the potential gains to the loser).

40. A point also made by Zerbe (2007).

41. For a more extended discussion of moral problems with both efficiency measures, see Coleman (1988, chapter 4).

42. Zerbe's (2001) approach thus goes against the standard economic analysis that assumes individuals are self-interested. To distinguish his approach from the standard approach, he refers to the standard approach as *cost-benefit analysis* and his alternative as *benefit-cost analysis*.

43. Euthyphro 10d1–8.

44. Cynics will say that these examples show only that businesses are willing to spend money to promote the image that they care about equity or environmental protection, not that they are willing to spend money to actually promote equity or environmental protection. However, even if the image is the goal, often the best way to promote the image of caring about equity and environmental protection is to actually promote them.

Chapter 10

1. For Scanlon, judgments of right and wrong "are judgments about what would be permitted by principles that could not reasonably be rejected, by people who were moved to find principles for the general regulation of behavior that others, similarly motivated, could not reasonably reject" (1998, 4).

2. Habermas's theory is only one of a number of ideal process approximation theories of deliberative democracy. A partial list of others would include J. Cohen (1989, 1997, and 1998); Gould (1988); Gutmann and Thompson (1996); and Young (2000). Although my criticisms of Habermas address the details of his account, I believe that all of these accounts are subject to at least some similar objections, but it would take me to far afield to address them all here.

3. "Thus the normative expectation of rational outcomes is grounded ultimately in the interplay between institutionally structured political will-formation and spontaneous, unsubverted circuits of communication in a public sphere that is not programmed to reach decisions and thus is not organized" (1996, 485).

4. What about other democracies? I don't have statistics. I would like to know how many democratic politicians have ever even lived in households with less than the national median income.

5. This is an instance of the truth that *any* decision rule can lead to abuse, even the rule of unanimity. Suppose, for example, that the settlers unanimously adopt a constitution that permits amending it with a 3/4 supermajority. At the time the constitution is adopted, the group is homogeneous, so there is no concern about a tyrannous majority. However, over time, one subgroup of 10% of the population becomes stigmatized and the other 90% amend the constitution to make slaves of the 10%. The fact that the slavery was adopted by unanimously agreed upon procedures would not justify it.

6. Because there is no reason to prefer one coin to another and the goal is just to single out one from a large number of equivalent alternatives, there is no danger of indeterminacy in majority rule for this kind of case, because there is no danger of intransitivity of the kind exploited in Arrow's (1963) impossibility theorem, in this kind of case.

7. This idea of a results-sensitive conception of equal respect is the main idea in Christiano's (2008) defense of a constitutional democracy. I should also mention that not even Waldron gives a purely procedural defense of majority rule, because he acknowledges that there are cases—"peculiar pathologies, dysfunctional legislative institutions, corrupt political cultures, legacies of racism and other forms of endemic prejudice" (2006, 1402)—in which judicial review is appropriate. I would only add that, typically, we

detect corrupt political cultures and some of the other items on Waldron's list by their results.

8. Fishkin heads the Center for Deliberative Democracy at Stanford. For more information, see the Center's Web site (http://cdd.stanford.edu/polls/docs/summary/).

9. Of course, election by deliberative poll is not the only way to solve this problem. Public financing of campaigns would do just as well, though it would probably be declared unconstitutional by the U.S. Supreme Court. See *Citizens United v. Federal Election Commission*, 558 U.S. 50 (2010). As Rawls argues persuasively (1993, 359–360), it is a serious mistake for the U.S. Supreme Court to place the freedom of wealthy people and corporations to buy political advertising above the importance of assuring fair elections.

10. For a fuller discussion of group rights grounded in individual autonomy rights, see Tan (2000).

11. *Citizens United v. Federal Election Commission*, 558 U.S. 50 (2010).

Chapter 11

1. The state of Michigan, the District of Columbia, and the cities of San Francisco, Palo Alto, and Santa Cruz have similar ordinances.

2. The main differences between us are that Nussbaum includes a central human capability "to be able to live with concern for and in relation to animals, plants, and the world of nature" (2000, 80). This seems to me to be an important value, but a different kind of value. Not all values are best understood as human capabilities. Nussbaum also includes a capability for play, which I agree is an important human capability, though not as important as most of the other items on her list.

3. This is not to say that there should be no repetitive, robotic minimum wage jobs full of drudgery. So long as everyone has reasonable alternatives, they should be free to choose such jobs, perhaps as a temporary way to make money to go to college.

4. Others who would concur that option luck does not raise issues of justice include G. A. Cohen (2008), and van Parijs (1991 and 1995). For a more sustained criticism of views of this kind, see Anderson (1999) and Hurley (2003).

5. In Talbott (1988), I use this rationale to explain why the courts favor cost-spreading agreements, but not benefit spreading agreements, in tort law.

6. I agree with Anderson's criticism of luck egalitarians, among whom she includes Dworkin, that it is a defect in their theories that they can justify making social insurance mandatory only on paternalistic grounds (1999, 301).

7. Note that this is not equivalent to maximin, because maximin would require reducing the income of the most well off by $1,000 to raise the income of the least well off by $1. However, whenever there are significant inequalities and there is a practice that, if adopted, would transfer an amount $M from the most well off to the least well off with little or no effects on motivation to engage in productive activity and little or no transaction costs

(including the costs of the implementation practice), then the main principle would almost surely favor it, because the main principle gives extra weight to the life prospects of the less well off.

8. Dworkin reports that 25% of Medicare expenses are for medical care in the last 6 months of life (2000, 314).

Chapter 12

1. This characterization of legal paternalism is narrower than Shiffrin's characterization of *paternalistic behavior* (2000, 215–218) but broader than Feinberg's (1986) category of laws that involve harm to self. Shiffrin's broader characterization is more useful as a characterization of the moral phenomenon of paternalistic behavior. Mine is intended to address a narrower question about the law. However, my notion is broader than Feinberg's category of laws involving harm to self, because it also covers some laws that Feinberg (1988) would characterize as harmless wrongdoing. Prohibitions of sodomy, which I discuss in the next chapter, fall into this category. In combining Feinberg's category of harm to self with at least some cases from his category of harmless wrongdoing, I am taking the Millian position that justifications for paternalistic laws typically characterize the prohibited conduct as "foolish, perverse, or wrong" ([1859], 19).

2. Of course, they could regard themselves as better off because other people are not engaging in the prohibited activity, so that they would not worry that those other people were harming themselves. Here I disregard any increase in well-being simply due to relief that others are not harming themselves, not because this sort of relief does not count as a contributor to well-being, but rather on the same grounds that Mill [1859] thought it should be disregarded, that the practice of disregarding it promotes well-being. Generally speaking, any increase in well-being due to such relief is more than outweighed by the gains in well-being from allowing people freedom to act on their autonomous judgments of what is good for them.

My characterization of the difference between legal solutions to CAPs and legal paternalism is not precise, because some paternalism involves a division in the self. In such cases, though paternalism overrules the judgment of the self at one time, it may give effect to the judgment of the very same self at a different time. The more precise characterization of CAPs is that they are situations in which, even if everyone in the relevant group acts rationally, the outcome may be worse for everyone in the group their own estimation than it would have been if they had all chosen differently (M. Taylor 1987, 19). Legal paternalism does not involve this kind of collective benefit.

3. I explain what I mean by *normal* adults in the next chapter. I discuss and defend the claim of first person authority in chapters 9 and 14 and in Talbott (2005, 123–128).

4. Mill ([1859], 109), quoted with approval by Feinberg (1986, 124). The names in brackets are my additions to Mill's example.

5. Of course, there is no guarantee that Feinberg would have accepted it as an improvement. Nothing crucial hinges on whether it is an improvement

or not, because the problems that I raise later in this chapter for the improved account are also problems for Feinberg's own account.

6. There seems to be a new movement to challenge Feinberg's voluntariness standard (e.g., Arneson 2005; Scoccia 2008; Shafer-Landau 2005). All these authors use examples involving drugs or suicide or both to raise problems for Feinberg's account. (Shafer-Landau [2005] uses the example of suicide to also raise questions about voluntary removal of limbs.) I take this movement as independent evidence that there is a problem with Feinberg's nonconsequentialist account. I can't tell whether my consequentialist version of the line between soft and hard paternalism would satisfy these authors, but it moves the line in the direction that they favor.

7. CNN (2001). Because this survey also shows that an overwhelming majority of smokers began smoking before age 18 (indeed, a majority began smoking before age 16), there is no basis for thinking their decision to begin smoking was an autonomous one. I set aside that issue here, because I am interested in the theoretical possibility that intervention could be justified even in cases in which a person makes an autonomous decision to begin drug use.

8. Although Nagel does not make explicit the assumption that both the earlier and later selves regard the other's judgment as less reliable than their own, I think this assumption is implicit in his discussion (1970, 74).

9. Of course, nonconsequentialists may insist that such precommitment devices should be enforceable. Thus, Feinberg (1986, 83), Nozick (1974, 331), G. Dworkin (1983, 111), and Thomson (1990, 283) all endorse enforcement of voluntary slavery contracts—at least in theory. Feinberg even discusses the Russian nobleman example and agrees that his earlier self should be able to bind his later self (1986, 86–87). To his credit, Rawls (1971) does not endorse enforcement of such contracts, but as I discussed in chapter 10, for the wrong reasons.

10. See Millgram (1997) for a much more detailed account of the extent to which practical reasoning involves learning, including learning about what is valuable or is worth desiring. Although I think most people are implicitly committed to the strong time-relative version of the claim of first-person authority most of the time, I do not insist that the evidence for its truth is decisive. As I explained in my discussion of the claim of first-person authority (Talbott 2005, 127), I believe that we are engaged in a long-term social experiment that will determine whether the claim is true.

11. It should be clear that my claim that people's later autonomous judgments about what is good for them are more reliable than their earlier ones is not an implicit argument for raising the voting age. That would replace the judgments of young people's earlier selves about what is good for them with the judgments of *other people's* later selves about what is good for them, not their own later selves. Also, as I explained in my first volume, I think democracies would be neither stable nor just if people always voted on the basis of their own good (Talbott 2005, chap. 7).

12. Regan (1983) considers this kind of justification of paternalism, but he does not draw any general conclusions from his discussion (113), so it is not clear whether he would endorse such a condition. In the course of his discussion, Regan also considers reasons for giving the later self's judgment

priority over the earlier self's judgment (129). So he was close to some kind of future endorsement standard, though a more modest one than mine, because he is more inclined to view the later self as a different person from the earlier one, as illustrated by his discussion of the example of the Russian nobleman (132–134).

13. When I state the probabilistic version of the most reliable judgment standard, it will only require that, in both branches of the diagram, there be a high probability of future endorsement (not necessarily that there be a high probability of bilateral future endorsement). See note 17 for an explanation of the difference.

14. Quinn Rotchford has suggested that in some cases, such as drug use, there is a basis for thinking that the judgments in one branch of the diagram are more reliable than those in the other branch. Rotchford thinks there would be general agreement that the judgments of the drug user would be more reliable than the judgments of the person who has been prevented from using the drug. If he is right, I don't see why this would be a problem for the most reliable judgment standard, because if there were general agreement that the judgments of the drug user would be more reliable than the judgments of the person prevented from using the drug, then even those who were prevented from using the drug would endorse the prohibition if those who used the drug did. In any case, my objection to Rotchford's proposal is that just as it makes sense to think that the person who uses the drug has a more reliable judgment about its effects, it makes sense to think that the person who is subject to the drug ban has a more reliable judgment about the effects of the ban. The future bilateral endorsement condition is a way of making sure that the hypothetical self who knows most about the drug and the hypothetical self who knows most about the effects of the ban both have a voice.

15. Because this is not a situation in which it is important to distinguish conditional probability from the probability of the corresponding subjunctive, I use the familiar symbol "/" (slash) for conditional probability in the text to represent the relevant subjunctive probability. For more on the distinction, see Gibbard and Harper (1978). Note also, as I explain in note 17, that this is not the same as the probability that one and the same individual would come to endorse intervention in both branches.

16. The argument in the text depends on the addition of one further qualification, which I include in the final statement of the most reliable judgment standard. The qualification is meant to address the following kind of case: If there is a sub-class of the target population that can be reliably distinguished at a reasonable cost from the other members of the target population for which Prob(UE/PI) and PROB(UE/-PI) are not both greater than .5, then that sub-class should be excluded from the target population.

17. Thanks to John Gresham for drawing my attention to the distinction between the future bilateral majority endorsement standard (the one I adopt in the text) and the future majority bilateral endorsement standard (which I do not adopt). As Gresham points out, it is possible that Prob(UE/PI) and Prob(UE/-PI) are both greater than 0.5 for the target population (the future bilateral majority endorsement standard is satisfied), but that the people in the target population who would come to endorse intervention given PI are,

for the most part, not the same as the people in the target population who would come to endorse intervention given -PI. Thus, the probability that a single individual member of the target population would come to endorse intervention in both branches of the diagram could be very low, much less than 0.5 (in which case, the future majority bilateral endorsement standard would not be satisfied). This shows that I must choose between the two majority conditions. The condition in the text, the future bilateral majority endorsement condition, is the appropriate condition for a consequentialist. The reason is the same reason as the reason for endorsing majority rule for legislation. When the future bilateral majority endorsement standard is satisfied, it is reasonable to believe that a majority of the target population will come to favor intervention in the case in which there is intervention and the case in which there is not (even if the two majorities will have different members). So majority rule favors intervention.

18. This is slightly different from the case of Dax Cowart, a burn victim who was kept alive despite his vigorous protests and who now has a life that is endorsed by his stable judgment about how to further his settled values and preferences, but still insists that he should not have been kept alive against his will when he was being treated for his burns (Childress and Campbell, 1997). If Cowart ever comes to the conclusion that, all things considered, it was better for him to have been kept alive against his will than to have been allowed to commit suicide, then his case will be an example of soft legal paternalism under the most reliable judgment standard (or it would be if those who had kept him alive had had good reason to expect that people in his situation would change their mind).

19. I should add that I do not mean to suggest the considerations raised here are the only reasons to oppose euthanasia. There are many reasons to oppose it. Because I am investigating the pure theory of legal paternalism, I have set those other considerations aside.

20. The most reliable judgment standard is based on an analogy with majoritarian democracy. In a majoritarian democracy, majority rule must be limited to cases in which its operation does not severely disadvantage a minority. So there must be some constraint that blocks majority will when it would seriously disadvantage a minority. Here I incorporate a similar constraint into the most reliable judgment standard.

21. This narrowest reference class requirement is a reflection of the fact that the expectations in the principle are statistical expectations. Statistical expectations are relative to a reference class. Let me give a simple example. Suppose that there is a drug RD on which 60% of potential users would come to unequivocally endorse a prohibition, both if there were a prohibition and if there were not. But now suppose that the target population of potential users is composed of two sub-populations of equal size, those with gene A and those with gene –A. Suppose that drug RD is much more addictive for those with gene A than for those with gene –A and so, only 40% of those with gene –A would come to endorse the prohibition, while 80% of those with gene A would come to support the prohibition.

Without this qualification, the most reliable judgment standard would classify a complete prohibition on drug RD as soft paternalism. However, if there were a simple and reliable means of distinguishing between those with

gene A and those with gene –A, the standard should not classify a general prohibition as soft paternalism. Only a prohibition on RD for those with gene A should be classified as soft paternalism. Finally, the degree of reliability of the means of distinguishing members of the sub-class and the reasonableness of the cost of doing so are contextual factors that can vary from case to case.

Chapter 13

1. I set aside here the discussion of other nonpaternalistic justifications that might be given for opposing a right to freedom of religion.

2. 381 U.S. 479 (1965).

3. The Court extended this right to unmarried people in *Eisenstadt v. Baird*, 405 U.S. 438 (1972).

4. 388 U.S. 1, 87 S.Ct. 1817 (1967).

5. 87 S.Ct. at 1824.

6. 505 U.S. 833, 851 (1992).

7. 478 U.S. 186 (1986).

8. *Powell v. State*, 510 S.E.2d 18 (1998).

9. 123 S.Ct. 2472 (2003).

10. 123 S.Ct. at 2481. By defining the right as a liberty right, the Court moved it from the penumbra and gave it a home in the Due Process Clause of the Fifth and Fourteenth Amendments.

11. The United States is no longer in the forefront of the development of a right against legal paternalism. A Canadian court has invalidated legal prohibitions on same-sex marriage and the legislature has acquiesced in the decision. *Halpern* v. *Toronto* (2003) 36 R.F.L. (5th) (Ontario Ct. App.). In Europe, almost all countries recognize same-sex unions, and Belgium, the Netherlands, Spain, Norway, and Sweden have legalized same-sex marriage. South Africa is the only other country to have recognized same-sex marriage, though many countries around the world are in the process of doing so.

12. *In re Quinlan*, 70 N.J. 10, 355 A.2d 647, cert. denied 429 U.S. 922 (1976).

13. *Cruzan v. Director, Missouri Dep't of Health*, 497 U.S. 261 (1990).

14. The Supreme Court's ruling came in the companion cases *Washington v. Glucksberg*, 117 S.Ct. 2258 (1997) and *Vacco v. Quill*, 117 S.Ct. 2293 (1997). For the Philosophers' Brief, see Dworkin, Nagel, Nozick, Rawls, Scanlon, and Thomson (1997).

15. *Gonzalez v. Oregon*, 546 U.S. 243 (2006).

16. For a discussion of a variety of other kinds of paternalistic government action that would be expected to satisfy any reasonable standard of soft paternalism, see Thaler and Sunstein (2008).

17. Unfortunately, we do not have statistical information that directly addresses this issue. However, indirect evidence indicates that the most reliable judgment standard is very likely satisfied. Social Security is one of the most popular government programs—especially among retirees. Feinberg claims that his voluntariness standard gives the same result, because Social Security is recognized by the majority to be in their interests, and enrollment is mandatory in order to achieve the economies of scale from covering

the entire population (1986, 18). This argument is unsatisfactory, because Feinberg's appeal to the majority view implicitly gets support from the large majority of retirees who favor Social Security. But for Feinberg's voluntariness standard, the crucial question would be: What are the opinions of young people who, if given the choice, would opt out or would not reliably make contributions if the deductions were not mandatory? They are the target population, because they are the only ones whose decisions are affected by the law's making the deduction mandatory. If early in their careers a majority of them do not support mandatory deductions, Social Security would be hard paternalism on Feinberg's view. Feinberg should have excluded the opinions of retirees and of those late in their careers, because, on the voluntariness standard, their opinions are irrelevant to whether or not the paternalism is soft or hard.

18. 123 S.Ct. 2472 (2003).

19. Of course, it is always possible that those who were restrained really didn't want to die and are not representative of those who actually jumped. These kinds of questions will always arise for the kind of statistical data we can get on suicides. However, we have good reason to believe that a high percentage of those who jumped to their deaths from the Golden Gate Bridge would also have changed their minds, because a small percentage of those who jump actually survive. When Tad Friend (2003) attempted to contact as many of the survivors as he could, he found that many of them also regretted jumping, some immediately after they leapt over the rail.

20. Feinberg tries to avoid the implication by suggesting other reasons for refusing to enforce slavery contracts (1986, 79–81). G. Dworkin just hopes the problem never comes up (1983, 111). Rawls's (1993) theory gives the right result in this case, but for the wrong reasons, as I explained in the discussion of inalienability of voting rights in chapter 10.

21. G. Dworkin advocates a formula of this kind: "Paternalism is justified only to preserve a wider range of freedom for the individual in question" (1972, 28).

22. The rule accepted in almost all jurisdictions is that a competent adult has the "basic right . . . to refuse treatment even when the treatment may be necessary to preserve the person's life." *In re Fosmire v. Nicoleau*, 551 N.E.2d 77, 81 (N.Y. 1990).

23. The religious cases introduce a complication into the analysis: Suppose that a Jehovah's Witness who was forced to undergo a transfusion thinks that it is wrong for him to *endorse* having been be forced to undergo a life-saving transfusion, but also believes that his life is the better for it. This kind of example requires me to reiterate what I mean by *endorsement* in the most reliable judgment standard. What is important is that the subject's judgment that the paternalistic intervention was good for her. No additional *act* of endorsement is required. This means that if Jehovah's Witnesses thought that it was wrong for them to truly answer a survey question that asked whether a blood transfusion made their life better, we might have no way of finding out that they really did endorse the transfusion. Another way to put this point is this: If some day we have mind-reading machines, we could use a mind-reading machine to find out whether Jehovah's Witnesses who have been forced to receive blood transfusions judge that their lives are the better for it

and, if so, that judgment, not any act of endorsement or failure to endorse, would be the evidence that would be relevant for applying the most reliable judgment standard.

There are other issues raised by the Jehovah's Witness example. Forced medical care in such a case infringes a right to religious freedom. This considerably complicates the question, because the right to religious freedom has its own consequentialist weight when evaluated by the main principle. So even if there were statistical evidence that supported a policy of forcing adults as well as children to undergo lifesaving medical treatment, I would expect the Supreme Court to allow an exception for conscientious objectors based on the right to religious freedom.

Also there is an incentives problem with forced medical care. If it is known that hospitals are authorized to force people to have medical care against their wishes, people who object to such care will tend to avoid hospitals. This is an important consideration in favor of prohibiting forced medical care, even if it could be justified as soft legal paternalism.

24. Judith Faller (2002) emphasizes that rights for those with mental illness need not be an all-or-nothing matter.

25. My discussion of privacy rights as human rights is indebted to A. Moore (2010), whose arguments for universality convinced me when I was unsure.

Chapter 14

1. Of course, pure utilitarians would not face collective action problems if they all agreed on all the relevant probabilities. Thus, omniscient pure utilitarians would have no need of moral practices. But their omniscience would make it easy for them to appreciate the importance of such practices if they had to interact with less-than-omniscient agents.

This page intentionally left blank

References

African Charter on Human Rights. 1986. In Center for the Study of Human Rights 1994: 119–128.

American Law Institute. 1977. *Restatement 2d of Torts* (St. Paul: American Law Institute Publishers).

———. 1981. *Restatement 2d of Contracts* (St. Paul: American Law Institute Publishers).

———.1985. *Model Penal Code* (Philadelphia).

Anderson, Elizabeth. 1993. *Value in Ethics and Economics* (Cambridge, Mass.: Harvard University Press).

———. 1999. What Is the Point of Equality? *Ethics* 109: 287–337.

Annan, Kofi. 2006. In-Depth Study on All Forms of Violence against Women: Report of the Secretary-General to the U.N. General Assembly. http://reliefweb.int/rw/lib.nsf/db900sid/HVAN-6UFSCZ/$file/UNGA-women-jul2006.pdf?openelement.

Apel, Karl-Otto. 1988. *Diskurs und Verantwortung* (Frankfurt: Suhrkamp).

Ariely, Dan. 2008. *Predictably Irrational* (New York: HarperCollins).

Aristotle. *Nicomachean Ethics*. In McKeon 1941: 935–1112.

———. *Politics*. In McKeon 1941: 1127–1316.

Arneson, Richard. 2005. Joel Feinberg and the Justification of Hard Paternalism. *Legal Theory* 11: 259–284.

Arrow, Kenneth J. 1963. *Social Choice and Individual Values* (New York: Wiley).

Barkow, Jerome H., Leda Cosmides, and John Tooby, eds. 1992. *The Adapted Mind* (New York: Oxford University Press).

Barry, Brian. 1995. *Justice as Impartiality* (Oxford: Clarendon Press).

Bennett, Jonathan. 1974. The Conscience of Huckleberry Finn. *Philosophy: The Journal of the Royal Institute of Philosophy* 49: 123–134.

Benson, Peter. 2001a. *The Theory of Contract Law: New Essays* (Cambridge: Cambridge University Press).

———. 2001b. The Unity of Contract Law. In Benson 2001a: 118–205.

Berlin, Isaiah. 1969. Two Concepts of Liberty. In *Four Essays on Liberty* (Oxford: Oxford University Press): 118–172.

Blake, Michael. 2001. Distributive Justice, State Coercion, and Autonomy. *Philosophy and Public Affairs* 30: 257–296.

———. Forthcoming. *Foreign Policy and Liberal Justice* (New York: Oxford University Press).

Boardman, William S. 1987. Coordination and the Moral Obligation to Obey the Law. *Ethics* 97: 546–557.

Bohman, James, and William Rehg, eds. 1997. *Deliberative Democracy* (Cambridge, Mass.: MIT Press).

Bosworth, Barry, Gary Burtless, and Claudia Sahm. 2001. The Trend in Lifetime Earnings Inequality and its Impact on the Distribution of Retirement Income, Center for Retirement Research at Boston College, Working Paper #2001-03. http://crr.bc.edu/working_papers/the_trend_in_lifetime_earnings_inequality_and_its_impact_on_the_distribution_of_retirement_income_3.html

Brandt, Richard B. 1992. *Morality, Utilitarianism, and Rights*. (New York: Cambridge University Press).

Brokaw, Tom. 1998. *The Greatest Generation* (New York: Random House).

Buchanan, Allen. 2004a. *Justice, legitimacy, and Self-Determination: Moral Foundations for International Law* (Oxford: Oxford University Press).

———. 2004b. Political Liberalism and Social Epistemology. *Philosophy and Public Affairs* 32: 95–130.

———. 2008. Human Rights and the Legitimacy of the International Order. *Legal Theory* 14: 39–70.

Burke, Edmund [1790] 1987. *Reflections on the Revolution in France*, J. G. A. Pocock, ed. (Indianapolis: Hackett).

Calabresi, Guido. 1970. *The Cost of Accidents* (New Haven, Conn.: Yale University Press).

Calhoun, Cheshire. 2009. What Good Is Commitment? *Ethics* 119: 613–641.

Card, David, and Alan B. Krueger. 1994. Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. *American Economic Review* 84: 772–793.

Center for the Study of Human Rights. 1994. *Twenty-Five Human Rights Documents* (New York: Columbia University).

Chang, Jung, and Jon Halliday. 2005. *The Unknown Story of Mao* (New York: Alfred A. Knopf).

Childress, James, with Courtney S. Campbell. 1997. Who Is a Doctor to Decide Whether a Person Lives or Dies? Reflections on Dax's Case. In James Childress, *Practical Reasoning in Bioethics* (Bloomington: Indiana University Press): 121–140.

Christiano, Thomas. 1997. The Significance of Public Deliberation. In Bohman and Rehg 1997: 243–78.

———. 2008. *The Constitution of Equality* (New York: Oxford University Press).

CNN. 2001. Gallup Poll: 82 Percent of U.S. Smokers Say They Would Like to Quit (January 5). http://www.tobaccofreedom.org/issues/cessation/survey_2001.html.

Cohen, G. A. 1995. *Self-Ownership, Freedom, and Equality* (Cambridge: Cambridge University Press).

———. 2008. *Rescuing Justice and Equality* (Cambridge, Mass.: Harvard University Press).

Cohen, Joshua. 1989. Deliberation and Democratic Legitimacy. In Alan Hamlin and Phillip Petit, eds., *The Good Polity: Normative Analysis of the State* (New York: Blackwell): 17–34.

———. 1997. Procedure and Substance in Deliberative Democracy. In Bohman and Rehg 1997: 407–437.

———. 1998. Democracy and Liberty. In Elster 1998: 185–231.
Coleman, Jules. 1988. *Markets, Morals, and the Law* (Cambridge: Cambridge University Press).
Confucius. 1995. *Analects*, William E. Soothill, trans. (New York: Dover).
Cosmides, Leda, and John Tooby. 1992. Cognitive Adaptations for Social Exchange. In Barkow, Cosmides, and Tooby 1992: 163–228.
Craswell, Ricard. 2001. Two Economic Theories of Enforcing Promises. In Benson 2001a: 19–44.
Curriden, Mark. 2000. Putting the Squeeze on Federal Juries. *American Bar Association Journal* 86: 52.
Dancy, Jonathan. 2004. *Ethics without Principles* (Oxford: Clarendon Press).
Darwall, Stephen. 2006. *The Second-Person Standpoint: Morality, Respect, and Accountability* (Cambridge, Mass.: Harvard University Press).
Dawes, R. M., Orbell, J. M., and Van de Kragt, J. C. 1986. Organizing Groups for Collective Action. *American Political Science Review* 80: 1171–1185.
Declaration of a Global Ethic. 1993. *1993 Parliament of the World's Religions* (Chicago). http://www.religioustolerance.org/parliame.htm.
Delaney, Neil Francis. 2007. A Note on Intention and the Doctrine of Double Effect. *Philosophical Studies* 134: 103–110.
DeNavas-Walt, Carmen, Bernadette D. Proctor, and Jessica C. Smith. 2008. *Income, Poverty, and Health Insurance Coverage in the United States: 2007*, U.S. Census Bureau, Current Population Reports, P60–235 (Washington, D.C.: U.S. Government Printing Office).
de Waal, Franz. 2006. *Primates and Philosophers: How Morality Evolved* (Princeton, N.J.: Princeton University Press).
Dhimgjoka, Merita. 2002. Blood Feuds Spark Call for Revised Killing Rules. *Seattle Times* (May 20): A8.
Donnelly, Jack. 2003. *Universal Human Rights in Theory and Practice*, 2nd ed. (Ithaca, N.Y.: Cornell University Press).
Dretske, Fred. 1971. Conclusive Reasons. *The Australasian Journal of Philosophy* 49: 1–22.
Dworkin, Gerald. 1972. Paternalism. *The Monist* 56: 64–84. Reprinted in Sartorius 1983: 19–34 [citations are to this version].
———. 1983. Paternalism: Some Second Thoughts. In Sartorius 1983: 105–111.
Dworkin, Ronald. 1977. *Taking Rights Seriously* (Cambridge, Mass.: Harvard University Press).
———. 1986. *Law's Empire* (Cambridge, Mass.: Belknap Press).
———. 1996. Objectivity and Truth: You'd Better Believe It. *Philosophy and Public Affairs* 25 (Spring): 87–139.
———. 2000. *Sovereign Virtue* (Cambridge, Mass.: Harvard University Press).
———. 2002. *Sovereign Virtue* Revisited. *Ethics* 113: 106–143.
Dworkin, Ronald, Thomas Nagel, Robert Nozick, John Rawls, Thomas Scanlon, and Judith Jarvis Thompson. 1997. Assisted Suicide: The Philosopher's Brief. *New York Review of Books* (March 27): 41–47.
Dwyer, William L. 2004. *In the Hands of the People: The Jury's Origins, Triumphs, Troubles, and Future in American Democracy* (New York: Thomas Dunne Books).

Elster, Jon. 1993. Majority Rule and Individual Rights. In Stephen Shute and Susan Hurley, eds., *On Human Rights: The Oxford Amnesty Lectures* (New York: Basic Books): 175–216.

———. 1998. *Deliberative Democracy* (Cambridge: Cambridge University Press).

Estlund, David M. 2008. *Democratic Authority* (Princeton, N.J.: Princeton University Press).

Faller, Judith L. 2002. *Who Qualifies for Rights*? (Ithaca, N.Y.: Cornell University Press).

Fehr, Ernst, and Simon Gächter. 2000. Cooperation and Punishment in Public Goods Experiments. *American Economic Review* 90: 980–994.

Feinberg, Joel. 1971. Legal Paternalism. *Canadian Journal of Philosophy* 1: 106–124. Reprinted in Sartorius 1983: 3–18 [citations are to this version].

———. 1980. Limits to the Free Expression of Opinion. *Philosophy of Law*, 2nd ed. (Belmont, Calif.: Wadsworth): 191–206.

———. 1986. *Harm to Self* (New York: Oxford University Press).

———. 1988. *Harmless Wrongdoing* (New York: Oxford University Press).

Fish, Stanley. 1994. *There's No Such Thing as Free Speech, and It's a Good Thing, Too* (New York: Oxford University Press).

Fishkin, James S. 1991. *Democracy and Deliberation: New Directions for Democratic Reform* (New Haven, Conn.: Yale University Press).

Fogel, Robert William, and Stanley L. Engerman. 1974. *Time on the Cross: The Economics of American Negro Slavery* (Boston: Little, Brown).

Frankfurt, Harry. 2000. The Moral Irrelevance of Equality. *Public Affairs Quarterly* 14: 87–103.

Freeman, Samuel. 2000. Deliberative Democracy: A Sympathetic Comment. *Philosophy & Public Affairs* 29: 371–418.

———. 2001. Illiberal Libertarians: Why Libertarianism Is Not a Liberal View, *Philosophy and Public Affairs* 30: 105–151.

———. 2007. *Justice and the Social Contract* (Oxford: Oxford University Press).

Fried, Charles. 1981. *Contract as Promise* (Cambridge, Mass.: Harvard University Press).

Friedman, Milton. 1962. *Capitalism and Freedom* (Chicago: University of Chicago Press).

———. 1970. The Social Responsibility of Business Is to Increase Its Profits. *New York Times Magazine* (September 13): 17.

Friend, Tad. 2003. Jumpers: The Fatal Grandeur of the Golden Gate Bridge. *New Yorker* (October 13): 48–59.

Gardiner, Stephen. Forthcoming. *The Perfect Moral Storm* (Oxford: Oxford University Press).

Garrett, Brandon L. 2008. Judging Innocence. *Columbia Law Review* 108: 101–190.

Gaus, Gerald. 1997. Reason, Justification, and Consensus: Why Democracy Can't Have It All. In Bohman and Rehg 1997: 205–242.

Gibbard, Allan. 1990. *Wise Choices, Apt Feelings* (Cambridge, Mass.: Harvard University Press).

Gibbard, Allan, and William L. Harper. 1978. Counterfactuals and Two Kinds of Expected Utility. In *Foundations and Applications of Decision Theory*, vol. I (Dordrecht: D. Reidel): 125–162.

Gilbert, Daniel. 2006. *Stumbling on Happiness* (New York: Alfred A. Knopf).
Giles, Jim. 2005. Internet Encyclopedias Go Head to Head. *Nature* 438 (December 15): 900–901.
Gilligan, Carol. 1982. *In a Different Voice* (Cambridge, Mass.: Harvard University Press).
Goldman, Alvin. 1976. Discrimination and Perceptual Knowledge. *The Journal of Philosophy* 73: 771–791.
Gould, Carol. 1988. *Rethinking Democracy: Freedom and Social Cooperation in Politics, Economy, and Society* (Cambridge: Cambridge University Press).
Greif, Avner. 2006. *Institutions and the Path to the Modern Economy*: *Lessons from Medieval Trade* (Cambridge: Cambridge University Press).
Griffin, James. 1988. *Well-Being: Its Meaning, Measurement and Moral Importance* (Oxford: Oxford University Press).
———. 2008. *On Human Rights* (Oxford: Oxford University Press).
Grisham, John. 2006. *The Innocent Man* (New York: Dell Trade Paperbacks).
Gutmann, Amy. 1987. *Democratic Education* (Princeton, N.J.: Princeton University Press).
Gutmann, Amy and Dennis Thompson. 1996. *Democracy and Disagreement* (Cambridge, Mass.: Harvard University Press).
Habermas, Jürgen. 1990. *Moral Consciousness and Communicative Action*, Christian Lenhardt and Shierry Weber Nicholsen, trans. (Cambridge, Mass.: MIT Press).
———. 1993. *Justification and Application*, Ciaran Cronin, trans. (Cambridge, Mass.: MIT Press).
———. 1995. Reconciliation through the Public Use of Reason: Remarks on John Rawls's Political Liberalism. *Journal of Philosophy* 92 (March): 109–131.
———. 1996. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*, William Rehg, trans. (Cambridge, Mass.: MIT Press).
———. 2003. *Truth and Justification*, Barbara Fultner, trans. (Cambridge, Mass.: MIT Press).
Halperin, Morton H., Joseph T. Siegle, and Michael M. Weinstein. 2005. *The Democracy Advantage: How Democracies Promote Prosperity and Peace* (New York: Routledge).
Hamilton, Edith, and Huntington Cairns. 1961. *The Collected Dialogues of Plato* (Princeton, N.J.: Princeton University Press).
Hampton, Jean. 1992. Rethinking Reason. *American Philosophical Quarterly* 29: 219–236.
Harman, Gilbert. 1977. *The Nature of Morality* (New York: Oxford University Press).
Harsanyi, John C. 1953. Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking. *Journal of Political Economy* 61: 434–435.
Hart, H. L. A. 1961. *The Concept of Law* (Oxford: Clarendon Press).
Hastings, Reed. 2009. Please Raise My Taxes. *New York Times* (February 6): A23.
Hayek, Friedrich. 1960. *The Constitution of Liberty* (Chicago: University of Chicago Press).

Hegel, Georg Wilhelm Friedrich. [1821] 1967. *Philosophy of Right*, T. M. Knox, trans. (New York: Oxford University Press).
Hobbes, Thomas. [1651] 1958. *Leviathan* (Indianapolis, Ind.: Library of Liberal Arts).
Hoffman, Abbie. 2002. *Steal This Book* (New York: Four Walls Eight Windows).
Hooker, Brad, and Margaret Olivia Little. 2000. *Moral Particularism* (Oxford: Clarendon Press).
Hugo, Victor. [1862] 1987. *Les Misérables*, Norman MacAfee, trans. (New York: Signet Classics).
Hume, David. [1740] 2005. *A Treatise of Human Nature*, David F. Norton and Mary J. Norton, eds. (Oxford: Oxford University Press).
———. [1777] 1970. *Enquiries Concerning the Human Understanding and Concerning the Principles of Morals*, L. A. Selby-Bigg, ed. (Oxford: Clarendon Press).
Hurley, Susan. 2003. *Justice, Luck, and Knowledge* (Cambridge, Mass.: Harvard University Press).
James, William. [1897] 2006. *The Will to Believe and Other Essays in Popular Philosophy* (New York: Cosimo).
Kahneman, Daniel, and Amos Tversky. 1979. Prospect Theory: An Analysis of Decision Under Risk. *Econometrica* 47: 263–292.
Kamm, F. M. 1989. Harming Some to Save Others. *Philosophical Studies* 57: 227–260.
Kant, Immanuel. [1785] 1959. *Foundations of the Metaphysics of Morals*, Lewis White Beck, trans. (New York: Liberal Arts Press).
———. [1793] 1974. *On the Old Saw: That May Be Right in Theory but It Won't Work in Practice*, E. B. Ashton, trans. (Philadelphia: University of Pennsylvania Press).
———. [1795] 1957. *Perpetual Peace*, Lewis White Beck, trans. (New York: Liberal Arts Press).
———. [1797] 1996. *The Metaphysics of Morals*, Mary Gregor, trans. and ed. (Cambridge: Cambridge University Press).
———. [1799] 1993. On a Supposed Right to Lie Because of Philanthropic Concerns, in *Grounding for the Metaphysics of Morals; With, On a Supposed Right to Lie Because of Philanthropic Concerns*, 3rd. ed., James W. Ellington, trans. (Indianapolis, Ind.: Hackett).
Keck, Thomas M. 2004. *The Most Activist Supreme Court in History: The Road to Modern Judicial Conservatism* (Chicago: University of Chicago Press).
Kim, Eunjung Katherine. 2009. *The Significance of an Overlapping Consensus on Human Rights* (Ph.D. dissertation: University of Washington).
Kitcher, Philip. 1990. The Division of Cognitive Labor. *The Journal of Philosophy* 87: 5–22.
Kohlberg, Lawrence. 1981. *The Philosophy of Moral Development* (San Francisco: Harper and Row).
Kornblith, Hilary. 1993. Epistemic Normativity. *Synthèse* 94: 357–376.
Korsgaard, Christine. 1986. Skepticism about Practical Reason. *The Journal of Philosophy* 83: 5–26.
———. 1996. *The Sources of Normativity* (New York: Cambridge University Press).

Kuhn, Thomas. 1962. *The Structure of Scientific Revolutions* (Chicago: University of Chicago Press).
Kymlicka, Will. 1989. *Liberalism, Community, and Culture* (Oxford: Clarendon Press).
LeBar, Mark. 2009. Virtue Ethics and Deontic Constraints. *Ethics* 119: 642–671.
Lafont, Cristina. 1998. Pluralism and Universalism in Discourse Ethics. In Amos Nascimento, ed., *A Matter of Discourse: Community and Communication* (Aldershot: Ashgate): 55–78.
Locke, John. [1690] 1952. *The Second Treatise of Government*, T. P. Peardon, ed. (Indianapolis, Ind.: Library of Liberal Arts).
MacIntyre, Alasdair. 1988. *Whose Justice? Which Rationality*? (Notre Dame, Ind.: University of Notre Dame Press).
Mackie, John. 1977. *Ethics: Inventing Right and Wrong* (New York: Penguin).
Marquis, Joshua. 2006. The Innocent and the Shammed. *New York Times* (January 26): A23.
Martin, J. Paul, and R. Rangaswamy, eds. 1994. *Twenty-Five Human Rights Documents* (New York: Columbia University Center for the Study of Human Rights).
Marx, Karl. [1867] 1982. Frederick Engels, ed., *Capital, Vol. 1*, in *Published Works, Vol. 35* (New York: Science & Society).
Marx, Karl and Friederich Engels. [1848] 1962. *The Communist Manifesto* (New York: Monthly Review Press).
Mayerfeld, Jamie. 2008. In Defense of the Absolute Prohibition of Torture. *Public Affairs Quarterly* 22: 109–128.
McDowell, J. 1978. Are Moral Requirements Hypothetical Imperatives? *Proceedings of the Aristotelian Society* 52: 13–29.
McKeon, Richard, trans. 1941. *The Basic Works of Aristotle* (New York: Random House).
Mele, Alfred R. 2001. *Self-Deception Unmasked* (Princeton, N.J.: Princeton University Press).
Mill, John Stuart. [1848] 1987. *Principles of Political Economy* (Fairfield, NJ: Augustus M. Kelley).
———. [1859] 1986. *On Liberty* (New York: Prometheus).
———. [1861] 1958. *Considerations on Representative Government* (New York: Liberal Arts Press).
———. [1863] 1965. *Utilitarianism*. In Schneewind 1965: 275–344.
———. [1873] 1924. *Autobiography of John Stuart Mill* (New York: Columbia University Press).
Millgram, Elijah. 1997. *Practical Induction* (Cambridge: Harvard University Press).
Mills, Charles W. 1997. *The Racial Contract* (Ithaca, N.Y.: Cornell University Press).
Moore, Adam D. 1997. A Lockean Theory of Intellectual Property. *Hamline Law Review* 21: 65–108.
———. 2010. *Privacy Rights: Moral and Legal Foundations* (University Park, Penn.: Pennsylvania State University Press).
Moore, G. E. 1903. *Principia Ethica* (Cambridge: Cambridge University Press).

Murphy, Liam. 2000. *Moral Demands in Nonideal Theory* (New York: Oxford University Press).
Murphy, Liam, and Thomas Nagel. 2002. *The Myth of Ownership* (New York: Oxford University Press).
Nagel, Thomas. 1970. *The Possibility of Altruism* (Oxford: Clarendon Press).
———. 1986. *The View from Nowhere* (New York: Oxford University Press).
———. 1991. *Equality and Partiality* (New York: Oxford University Press).
———. 2005. The Problem of Global Justice. *Philosophy & Public Affairs* 33: 113–147.
Noddings, N. 1984. *Caring: A Feminine Approach to Ethics and Moral Education* (Berkeley, Calif.: University of California Press).
Nozick, Robert. 1974. *Anarchy, State, and Utopia* (New York: Basic Books).
———. 1981. *Philosophical Explanations* (Cambridge, Mass.: Belknap Press).
———. 1989. *The Examined Life* (New York: Simon and Schuster).
———. 1993. *The Nature of Rationality* (Princeton, N.J.: Princeton University Press).
Nussbaum, Martha C. 2000. *Women and Human Development* (Cambridge: Cambridge University Press).
Olson, Mancur. 2000. *Power and Prosperity* (New York: Basic Books).
Parfit, Derek. 1984. *Reasons and Persons* (Oxford: Clarendon Press).
———. 1997. Equality and Priority. *Ratio* 10: 202–221.
Peirce, Charles S. 1992/1998. *The Essential Peirce*, 2 vols., Nathan Houser, Christian Kloesel, and the Peirce Edition Project, eds. (Bloomington: Indiana University Press).
Pettit, Philip. 1997. *Republicanism: A Theory of Freedom and Government* (Oxford: Clarendon Press).
———. 2006. No Testimonial Route to Consensus. *Episteme* 3: 156–165.
Plato. *Crito*, Hugh Tredennick, trans. In Hamilton and Cairns 1961: 27–39.
———. *Euthyphro*, Lane Cooper, trans. In Hamilton and Cairns 1961: 169–185.
———. *Republic*, Paul Shorey, trans. In Hamilton and Cairns 1961: 575–844.
Pogge, Thomas W. 2000. The International Significance of Human Rights. *Journal of Ethics* 4: 45–69.
———. 2002. *World Poverty and Human Rights* (Cambridge, U.K.: Polity Press).
Posner, Richard A. 1983. *The Economics of Justice* (Cambridge, Mass.: Harvard University Press).
———. 2007. *The Economic Analysis of Law*, 7th ed. (New York: Aspen Law & Business).
Priest, Graham. 2006. *In Contradiction, A Study of the Transconsistent (Expanded Edition)* (Oxford: Clarendon Press).
Quinn, Warren. 1989. Actions, Intentions, and Consequences: The Doctrine of Double Effect. *Philosophy and Public Affairs* 18: 334–351; reprinted in Woodward 2001: 23–40.
Railton, Peter. 1984. Alienation, Consequentialism, and the Demands of Morality. *Philosophy and Public Affairs* 13: 134–171.
Rawls, John. 1955. Two Concepts of Rules. *Philosophical Review* 64: 3–32.
———. 1971. *A Theory of Justice* (Cambridge, Mass.: Harvard University Press).

———. [1975] 1999. A Kantian Conception of Equality. In Samuel Freeman, ed., *Collected Papers* (Cambridge, Mass.: Harvard University Press).
———. 1985. Justice as Fairness: Political not Metaphysical. *Philosophy & Public Affairs* 14: 223–252.
———. 1993. *Political Liberalism* (New York: Columbia University Press).
———. 1994. *Political Liberalism*, paperback edition (New York: Columbia University Press).
———. 1995. Reply to Habermas. *The Journal of Philosophy* 92: 132–180.
———. 1999. *The Law of Peoples* (Cambridge, Mass.: Harvard University Press).
———. 2001. *Justice as Fairness: A Restatement*, Erin Kelly, ed. (Cambridge, Mass.: Belknap Press).
Raz, Joseph. 1986. *The Morality of Freedom* (Oxford: Clarendon Press).
Regan, Donald. 1980. *Utilitarianism and Co-operation* (Oxford: Clarendon Press).
———. 1983. Freedom, Identity, and Commitment. In Sartorius 1983: 113-138.
Reidy, David. 2005. An Internationalist Conception of Human Rights. *Philosophical Forum* 36: 367–397.
———. 2008. William Talbott's Which Rights Should Be Universal? *Human Rights Review* 9: 181–191.
Risinger, D. Michael. 2007. Innocents Convicted: An Empirically Justified Factual Wrongful Conviction Rate. *Journal of Criminal Law and Criminology* 97: 761–804.
Ross, W. D. 1930. *The Right and the Good* (Oxford: Clarendon Press).
Rousseau, Jean-Jacques. [1762] 1950. *The Social Contract, and Discourses* (New York: Dutton).
Ruse, Michael. 1995. Evolutionary Ethics: A Phoenix Arisen. In P. Thompson, ed., *Issues in Evolutionary Ethics* (Albany: State University Press of New York): 225-248.
Sartorius, Rolf, ed. 1983. *Paternalism* (Minneapolis: University of Minnesota Press).
Satz, Debra. 2009. Voluntary Slavery and the Limits of the Market. *Law & Ethics of Human Rights* 3: 86–109.
Scanlon, T. M. 1972. A Theory of Freedom of Expression. *Philosophy and Public Affairs* 1: 204–226.
———. 1975. Preference and Urgency. *Journal of Philosophy* 38: 91–109.
———. 1982. Contractualism and Utilitarianism. In A. Sen and B. Williams, eds., *Utilitarianism and Beyond* (Cambridge: Cambridge University Press): 103–128.
———. 1998. *What We Owe to Each Other* (Cambridge, Mass.: Belknap Press).
Schneewind, J. B., ed. 1965. *Mill's Ethical Writings* (New York: Collier Books).
Scoccia, Danny. 2008. In Defense of Hard Paternalism. *Law and Philosophy* 27: 351–381.
Seiden, Richard. 1978. Where Are They Now? A Follow-Up Study of Suicide Attempters from the Golden Gate Bridge. *Suicide and Life Threatening Behavior* 8: 203–16.

Sen, Amartya. 1999. *Development as Freedom* (New York: Alfred A. Knopf).
———. 2000. Consequential Evaluation and Practical Reason. *Journal of Philosophy* 97: 477–502.
———. 2003. *Rationality and Freedom*. (Cambridge, Mass.: Harvard University Press).
———. 2006. *Identity and Violence: The Illusion of Destiny* (New York: Norton).
———. 2009. *The Idea of Justice* (Cambridge, Mass.: Belknap Press).
Shafer-Landau, Ross. 2005. Liberalism and Paternalism. *Legal Theory* 11: 169–191.
Shiffrin, Seana. 2000. Paternalism, Unconscionability Doctrine, and Accommodation. *Philosophy & Public Affairs* 29: 205–250.
Shue, Henry. 1980. *Basic Rights* (Princeton, N.J.: Princeton University Press).
Skyrms, Brian. 1996. *Evolution of the Social Contract* (Cambridge: Cambridge University Press).
Smith, Adam. [1776] 1976. *An Inquiry into the Nature and Causes of the Wealth of Nations*, W. B. Todd, ed. (Oxford: Clarendon Press).
Sophocles. 1988. *Oedipus the King*, Stephen Berg, trans. (New York: Oxford University Press).
Stannard, David E. 1992. *American Holocaust* (New York: Oxford University Press).
Stich, Stephen. 1990. *The Fragmentation of Reason* (Cambridge, Mass.: MIT Press).
Styron, William. 1979. *Sophie's Choice* (New York: Random House).
Sunstein, Cass R. 2006a. Deliberating Groups versus Prediction Markets (or Hayek's Challenge to Habermas). *Episteme* 3: 192–213.
———. 2006b. *Infotopia: How Many Minds Aggregate Knowledge* (New York: Oxford University Press).
Surowiecki, James. 2004. *The Wisdom of Crowds* (New York: Doubleday).
Talbott, William J. 1988. Cost Spreading and Benefit Spreading in Tort Law. *Research in Law and Economics* 11: 25–51.
———. 1995. Intentional Self-Deception in a Single, Coherent Self. *Philosophy and Phenomenological Research* 55: 27–74.
———. 1998. Why We Need A Moral Equilibrium Theory. In Peter A. Danielson, ed., *Modeling Rationality, Morality, and Evolution* (Oxford: Oxford University Press): 302–339.
———. 2005. *Which Rights Should Be Universal*? (New York: Oxford University Press).
———. 2007. The Universality of Human Rights: A Response. *Human Rights and Human Welfare: An International Review of Books and Other Publications* 7: 113–141.
———. 2008. Reply to Critics: In Defense of One Kind of Epistemically Modest but Metaphysically Immodest Liberalism. *Human Rights Review* 9: 193–212.
Tan, Kok-Chor. 2000. *Toleration, Diversity, and Global Justice* (University Park: Pennsylvania State University Press).
Taylor, Charles. 1999. Conditions on an Unforced Consensus on Human Rights. In Joanne R. Bauer and Daniel A. Bell, eds., *The East Asian*

Challenge for Human Rights (New York: Cambridge University Press): 124–144.
Taylor, Michael. 1987. *The Possibility of Cooperation* (Cambridge: Cambridge University Press).
Temkin, Larry S. 1993. *Inequality* (Oxford: Oxford University Press).
Tetlock, Philip. 2005. *Expert Political Judgment: How Good Is It? How Can We Know*? (Princeton, N.J.: Princeton University Press).
Thaler, Richard H., and Cass R. Sunstein. 2008. *Nudge: Improving Decisions about Health, Wealth, and Happiness* (New Haven, Conn.: Yale University Press).
Thomson, Judith Jarvis. 1990. *The Realm of Rights* (Cambridge, Mass.: Harvard University Press).
Twain, Mark. [1884] 1981. *The Adventures of Huckleberry Finn* (New York: Bantam).
U.N. 1948. *Universal Declaration of Human Rights*. In Center for the Study of Human Rights 1994: 6–9.
———. 2009. Convention on the Rights of Persons with Disabilities. http://www.un.org/disabilities/default.asp?navid=12&pid=150.
van Parijs, Phillipe. 1991. Why Surfers Should Be Fed: The Liberal Case for an Unconditional Basic Income. *Philosophy & Public Affairs* 20: 101–131.
———. 1995. *Real Freedom for All: What (if Anything) Can Justify Capitalism*? (Oxford: Oxford University Press).
Veatch, Robert. 1987. *The Patient as Partner* (Bloomington: Indiana University Press).
von Platz, Jeppe. 2008. Reasonable Disagreement and Metaphysical Immodesty: A Comment on Talbott's *Which Rights Should Be Universal*? *Human Rights Review* 9: 167–179.
Waldfogel, Joel. 1993. The Deadweight Loss of Christmas. *American Economic Review* 83: 1328–1336.
Waldron, Jeremy. 1998. *Law and Disagreement* (Oxford: Oxford University Press).
———. 2006. The Core of the Case against Judicial Review. *Yale Law Journal* 115: 1346–1402.
Weart, Spencer R. 1998. *Never at War: Why Democracies Will Not Fight One Another* (New Haven, Conn.: Yale University Press).
Williams, Bernard. 1973. A Critique of Utilitarianism. In Bernard Williams and J. J. C. Smart, *Utilitarianism: For and Against* (Cambridge: Cambridge University Press): 77–150.
———. 1985. *Ethics and the Limits of Philosophy* (Cambridge, Mass.: Harvard University Press).
Wilson, E. O. 1998. *Consilience: The Unity of Knowledge* (New York: Knopf).
Woodward, P. A., ed. 2001. *The Doctrine of Double Effect: Philosophers Debate a Controversial Moral Principle* (Notre Dame, Ind.: University of Notre Dame Press.
Wright, Robert. 2000. *Nonzero: The Logic of Human Destiny* (New York: Pantheon Books).
Young, Iris Marion. 2000. *Inclusion and Democracy* (Oxford: Oxford University Press).

Zerbe, Richard O., Jr. 2001. *Economic Efficiency in Law and Economics* (Northampton, Mass.: Edward Elgar).
———. 2007. The Legal Foundation of Cost-Benefit Analysis, *Charleston Law Review* 2: 93–184.
Zerbe, Richard O., Jr., and C. Leigh Anderson. 2001. Culture and Fairness in the Development of Institutions in the California Gold Fields. *The Journal of Economic History* 61: 114–143.

Table of Cases

Bowers v. Hardwick, 478 U.S. 186 (1986).
Brandenburg v. Ohio, 395 U.S. 444 (1969).
Citizens United v. Federal Election Commission, 558 U.S. 50 (2010)
Cruzan v. Director, Missouri Dep't of Health, 497 U.S. 261 (1990).
Dennis v. United States, 341 U.S. 494 (1951).
Eisenstadt v. Baird, 405 U.S. 438 (1972).
Gonzalez v. Oregon, 546 U.S. 243 (2006).
Greenman v. Yuba Power Products, Inc., 59 Cal.2d 57 (1963).
Griffith v. Kentucky, 479 U.S. 314 (1987).
Griswold v. Connecticut, 381 U.S. 479 (1965).
Halpern v. *Toronto*,(2003) 36 R.F.L. (5th) (Ontario Ct. App.).
Henningsen v. Bloomfield Motors, Inc. and Chrysler Corporation., 32 N.J. 358; 161 A.2d 69 (1960).
In re Fosmire v. Nicoleau, 551 N.E.2d 77, 81 (N.Y. 1990).
In re Quinlan, 70 N.J. 10, 355 A.2d 647, cert. denied 429 U.S. 922 (1976).
Johnson v. New Jersey, 384 U.S. 719 (1966).
Kansas v. Marsh, 126 S. Ct. 2516, 165 L.Ed. 429 (2006).
Lawrence v. Texas, 123 S.Ct. 2472 (2003).
Loving v. Virginia, 388 U.S. 1, 87 S.Ct. 1817 (1967).
MacPherson v. Buick Motor Co., 217 N.Y. 382, 111 N.E. 1050 (1916).
Mapp v. Ohio, 367 U.S. 643 (1961).
Masses Publishing Co. v. Patten 245 F. 535 (1917).
Miranda v. Arizona, 384 U.S. 436 (1966).
New York Times v. Sullivan, 376 U.S. 254 (1964).
Planned Parenthood v. Casey, 505 U.S. 833, 851 (1992).
Powell v. State, 510 S.E.2d 18 (1998).
Regina v. Dudley and Stevens, 14 QBD 273 DC (1884).
Riggs v. Palmer, 115 N.Y. 506, 22 N.E. 188 (1889).
Roe v. Wade, 410 U.S. 113 (1973).
Rylands v. Fletcher, LR 3 HL 330 (1868).
Vacco v. Quill, 117 S.Ct. 2293 (1997).
Washington v. Glucksberg, 117 S.Ct. 2258 (1997).
West Coast Hotel Co. v. Parrish, 300 U.S. 379 (1937).
Williams v. Walker-Thomas Furniture Co., 350 F.2d 445 (D.C. Cir. 1965).

Index